Statistical Science
1992, Vol. 7, No. 1, 131–177

A Survey of Exact Inference for Contingency Tables

Alan Agresti

Abstract. The past decade has seen substantial research on exact inference for contingency tables, both in terms of developing new analyses and developing efficient algorithms for computations. Coupled with concomitant improvements in computer power, this research has resulted in a greater variety of exact procedures becoming feasible for practical use and a considerable increase in the size of data sets to which the procedures can be applied. For some basic analyses of contingency tables, it is unnecessary to use large-sample approximations to sampling distributions when their adequacy is in doubt. This article surveys the current theoretical and computational developments of exact methods for contingency tables. Primary attention is given to the exact conditional approach, which eliminates nuisance parameters by conditioning on their sufficient statistics. The presentation of various exact inferences is unified by expressing them in terms of parameters and their sufficient statistics in loglinear models. Exact approaches for many inferences are not yet addressed in the literature, particularly for multidimensional contingency tables, and this article also suggests additional research for the next decade that would make exact methods yet more widely applicable.

Key words and phrases: Categorical data, chi-squared tests, computational algorithms, conditional inference, Fisher's exact test, logistic regression, loglinear models, odds ratios, sufficient statistics.

Table of Contents

Alan Agresti is Professor of Statistics, University of Florida, Griffin-Floyd Hall, Gainesville, Florida 32611.

1. INTRODUCTION

This article surveys the development of exact inferential methods for contingency tables. I interrelate various inferences by expressing them in terms of parameters in a hierarchy of loglinear models. The presentation focuses primarily on exact conditional methods, in which one obtains sampling distributions not dependent on other unknown parameters by conditioning on their

sufficient statistics. I also discuss some long-standing controversies for such methods and present topics for additional research that should soon be feasible given the continual improvement in computer hardware and software.

1.1 Historical Perspective

Traditionally, statistical inference for contingency tables has relied heavily on large-sample approximations for sampling distributions of parameter estimators and test statistics. Many of these approximations are special cases of ones that apply more generally than to categorical data [e.g., chi-squared approximations for likelihood-ratio statistics and normal approximations for maximum likelihood (ML) estimators of model parameters]. With this emphasis on large-sample methods, the development of inferential methods for categorical data parallels the historically earlier development of inferential methods for continuous data. For instance, E. S. Pearson's recently published manuscript about Student notes that studies analyzed by Karl Pearson's laboratory usually involved large-size data sets. When Gosset presented his queries about small samples that led to his development of techniques using the t-distribution, Pearson replied, "Only naughty brewers deal in small samples" (Pearson, 1990, page 73).

R. A. Fisher's *Statistical Methods for Research Workers* was at the forefront of advocating exact procedures for small samples. In the preface of the first edition of that book (1925), Fisher stated, "... the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." Not surprisingly, t-procedures received strong emphasis in that text, and "Fisher's exact test" for 2×2 contingency tables appeared in the 1934 and subsequent editions.

The importance of improving the scope of exact methods for categorical data has become increasingly clear in recent years. Standard asymptotic methods apply to a fixed number of cells, as cell expected frequencies grow to infinity. Yet, researchers often attempt to analyze additional variables as the sample size grows; thus, large expected frequencies may be the exception rather than the norm. Although recent research has introduced new asymptotic approaches that permit the number of cells to grow as the sample size grows (e.g., Morris, 1975; Haberman, 1977; Koehler, 1986; McCullagh, 1986; Zelterman, 1987), information on the adequacy of these approximations for standard models is at an infant stage. [Cressie and Read (1989) surveyed research on the adequacy of various asymptotic approximations.] Also, simulation studies have shown that it is hopeless to expect simple guidelines to indicate when asymptotic large-sample approximations are adequate (e.g., Koehler and Larntz, 1980). Even when the sample size is quite large, recent work has shown that large-sample approximations can be very poor when the contingency table contains both small and large expected frequencies (Haberman, 1988). Regarding adequacy of asymptotics, the sample size n often has less relevance than the discreteness of the sampling distribution. Thus, "small-sample methods" for categorical data more accurately refer to methods needed when there are few points or relatively large probabilities in the support of that distribution.

Table 1, taken from Graubard and Korn (1987), illustrates that different statistics and approximations can give quite different results, even for very large samples. That table, which refers to a prospective study of maternal drinking and congenital malformations, has 32,574 observations. For testing independence of alcohol consumption and malformation, asymptotic chi-squared tests give p-values of 0.017 using the Pearson statistic and 0.190 using the likelihood-ratio statistic. Exact tests of a type discussed in Section 3.1 using these criteria give p-values of 0.034 and 0.139, respectively. A test based on a trend alternative that utilizes the ordering of column categories by assigning scores $(0, 0.5, 1.5, 4, 7)$ to them has exact p-value more than three times the p-value based on an asymptotic normal approximation (0.017 versus 0.005 in the one-sided case).

The lag in the development and use of exact inferences for contingency tables is partly explained by the later development of methods for categorical data compared with continuous data, but also by the greater computational complexity. However, recent advances in computational power

TABLE 1
Maternal drinking and congenital malformations

	Alcohol consumption (average no. of drinks/day)				
Malformation	0	< 1	1-2	3-5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Source: Graubard and Korn (1987).

and efficiency of algorithms have made exact methods feasible for a wider variety of inferential analyses and for a larger collection of table sizes and sample sizes. In this survey of exact inferences for contingency tables, I indicate which computations can currently be done and also highlight areas in which current capabilities are inadequate.

1.2 Outline and Notation

Sections 2 to 4 present a variety of exact methods for contingency tables in the context of loglinear models. Sections 2 and 3 focus on two-way tables, Section 2 for the 2×2 case and Section 3 for the $I \times J$ case. Section 4 focuses on three-way tables, with emphasis on the $2 \times 2 \times K$ case. Section 5 discusses exact inference for logistic regression models, and Section 6 discusses exact goodness-of-fit testing for loglinear and logistic models. Section 7 discusses computing feasibility of exact methods, using currently available software.

Nearly all the literature on exact methods for contingency tables emphasizes hypothesis testing; but, in each section, I also indicate the scope of work on interval estimation. My discussion throughout the article takes the viewpoint of classical frequentist conditional inference, with application to loglinear models for contingency tables. Section 8 mentions other approaches. It presents an unconditional approach to exact inference and also indicates how Bayesian inferences correspond to conditional inferences for certain choices of prior distributions. The final section discusses possible future directions for research on exact methods for contingency tables.

Throughout the article, I assume a standard Poisson or multinomial sampling model for cell counts in the contingency table. For instance, in an $I \times J$ table, the cell counts $\{n_{ij}\}$ might have a multinomial distribution generated by n independent trials with IJ cell probabilities $\{\pi_{ij}\}$. Or, the counts $\{n_{ij}, j = 1, \ldots, J\}$ in row i might have a multinomial distribution, with counts in different rows being independent. In the first case ("full" multinomial sampling), $n = \sum\sum n_{ij}$ is fixed. In the second case (independent multinomial sampling), $\{n_{i+} = \sum_j n_{ij}, i = 1, \ldots, I\}$ are fixed. When $\{n_{ij}\}$ are independent Poisson random variables, conditioning on n yields full multinomial sampling; conditioning further on the row totals yields independent multinomial sampling. All sampling models lead to the same exact inferences, since those inferences condition on marginal totals that contain as a subset the naturally fixed totals, and since the parameters of usual interest are not the proportions in the margins that are fixed under some sampling designs but not under others.

For two-way contingency tables, let X denote the row classification and Y the column classification. For three-way tables, denote the third classification by Z. For simplicity, denote loglinear models by standard symbols pertaining to their minimal sufficient statistics. For instance, (X, Y) denotes the model of statistical independence in a two-way table, and (XZ, YZ) denotes the model of conditional independence between X and Y, given Z, in a three-way table. The minimal sufficient statistics are $\{n_{i+}\}$ and $\{n_{+j}\}$ for (X, Y) and $\{n_{i+k}\}$ and $\{n_{+jk}\}$ for (XZ, YZ). Finally, denote expected frequencies by m and sample proportions by p, for example, $\{m_{ij} = E(n_{ij})\}$ and $\{p_{ij} = n_{ij}/n\}$ in a two-way table.

1.3 The Exact Conditional Approach

Historically, the most common approach to exact inference in contingency tables has been a conditional one. Suppose an inference refers to a parameter in some loglinear model. Exact conditional inferential methods utilize the distribution of the sufficient statistic for that parameter, conditional on sufficient statistics for the other parameters ("nuisance" parameters) in the model. For instance, suppose one wants to test a hypothesis H_0 that corresponds to a loglinear model symbolized by M_0, under the assumption that a more general model M_1 holds, corresponding to an alternative hypothesis H_1. Denote minimal sufficient statistics for the models by T_0 and T_1. Exact inference uses the conditional distribution of T_1 given T_0 (e.g., Andersen, 1974). By the definition of sufficiency, the conditional distribution does not depend on the nuisance parameters, thus making exact inference possible.

To illustrate, for Poisson sampling in a 2×2 contingency table, the saturated loglinear model has the form

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}. \tag{1.1}$$

The parameters $\{\lambda_{ij}\}$ describing association in this model pertain to the odds ratio. For instance, suppose we achieve identifiability by setting $\lambda_2^X = \lambda_2^Y = \lambda_{21} = \lambda_{12} = \lambda_{22} = 0$. Then $\lambda_{11} = \log(\theta)$, where θ denotes the odds ratio, $\theta = (m_{11}m_{22})/(m_{12}m_{21})$. The parameter set is $(\mu, \lambda_1^X, \lambda_1^Y, \lambda_{11})$, and normally interest focuses on $\theta = \exp(\lambda_{11})$, the others being nuisance parameters. The sufficient statistics are n for μ, n_{1+} for λ_1^X, n_{+1} for λ_1^Y, and n_{11} for λ_{11}. To conduct exact inference about θ, consider the conditional distribution of n_{11} given n, n_{1+}, and n_{+1}. This conditional distribution is the same as the one having elements $P(Y_1 = n_{11}, Y_2 = n_{+1} - n_{11} \mid Y_1 + Y_2 = n_{+1})$, where $\{Y_i, i = 1, 2\}$ are indepen-

dent binomial random variables with parameters $[n_{i+}, m_{i1}/(m_{i1} + m_{i2})]$. Straightforward calculation shows that it equals

$$f(n_{11} \mid n, n_{1+}, n_{+1}; \theta) = \frac{\binom{n_{1+}}{n_{11}}\binom{n - n_{1+}}{n_{+1} - n_{11}}\theta^{n_{11}}}{\sum_u \binom{n_{1+}}{u}\binom{n - n_{1+}}{n_{+1} - u}\theta^u}, \tag{1.2}$$

where the index of summation ranges from $\max(0, n_{1+} + n_{+1} - n)$ to $\min(n_{1+}, n_{+1})$, the possible values for n_{11} for the given marginal totals. This is the noncentral hypergeometric distribution (Fisher, 1935a; Cornfield, 1956).

To test statistical independence of X and Y [$\lambda_{11} = 0$ in model (1.1)], one uses this distribution with $\theta = 1$. Of course, to complete a test, one needs to specify a test statistic and the way to compute the p-value. In testing a loglinear model M_0 against a more complex model M_1, in this article I will generally base p-values on the exact conditional distribution of Rao's efficient score statistic for testing that the extra parameters that are in M_1 but not in M_0 equal zero. The efficient score statistic is based on the vector of partial derivatives of the log likelihood with respect to the extra parameters, evaluated at the null estimates (Rao, 1973, Section 6e). For testing independence in a 2×2 table, this suggests $[n_{11} - (n_{1+}n_{+1})/n]$, suitably normalized, as a test statistic.

The p-value, based on extreme values of the test statistic, is calculated using the distribution for that statistic that is induced by the exact conditional distribution of $\{n_{ij}\}$. When M_1 contains one more parameter than M_0, tests based on the efficient score statistic are uniformly most powerful unbiased (UMPU), as a consequence of a result for exponential families (e.g., Lehmann, 1986, Theorem 3, page 147). Denote the ML estimator of the parameters in M_1 by $(\hat{\theta}_0, \hat{\theta}_1)$, where $\hat{\theta}_1$ denotes the estimator of the parameters that are in M_1 but not in M_0. For large samples with a fixed number of cells, these tests are asymptotically equivalent to Wald tests based on $\hat{\theta}_1$ and likelihood-ratio tests of M_0 against M_1 (Rao, 1973, pages 418–420).

My reason for relating exact conditional inference to *loglinear* models in this article is as follows. Loglinear models are generalized linear models for categorical data that use the canonical link, that is, they directly model the natural parameter (the log mean) of a natural exponential family (the Poisson). Such generalized linear models permit reduction of the data through sufficient statistics (McCullagh and Nelder, 1989, page 32). Thus, one can eliminate nuisance parameters in the model by conditioning on sufficient statistics for them.

2. EXACT INFERENCE FOR 2 × 2 TABLES

This simple case still generates an enormous volume of publications, partly because of controversy (discussed in Section 8.1) about applying exact methods that condition on marginal totals that are not naturally fixed under Poisson, multinomial, or binomial sampling schemes. The most common sampling model for 2×2 tables is independent binomial sampling in the two rows. Denote the "success" probabilities by π_1 and π_2, for which the odds ratio is $\theta = [\pi_1/(1 - \pi_1)]/[\pi_2/(1 - \pi_2)]$. The hypothesis of homogeneity states that $\pi_1 = \pi_2$. For Poisson or full multinomial sampling, conditioning on $\{n_{1+}, n_{2+}\}$ gives binomial sampling, and statistical independence is equivalent to homogeneity. Under the hypothesis of homogeneity, conditioning as well on n_{+1} to eliminate the nuisance parameter (the common value of π_1 and π_2) yields the hypergeometric distribution (1.2) with $\theta = 1$.

2.1 Fisher's Exact Test

Fisher's exact test (Fisher, 1934, 1935a; Yates, 1934; Irwin, 1935) of H_0: $\theta = 1$ is "well known," and I refer the reader to Lehmann (1986, pages 151–162) for details of its derivation under various sampling schemes. The null hypergeometric distribution has $E(n_{11}) = n_{1+}n_{+1}/n$ and $\mathrm{Var}(n_{11}) = n_{1+}n_{+1}n_{2+}n_{+2}/n^2(n - 1)$. To test H_0: $\theta = 1$ against H_1: $\theta > 1$, the p-value is

$$P = \sum_S f(t \mid n, n_{1+}, n_{+1}; 1), \tag{2.1}$$

where $S = \{t: t \geq n_{11}\}$. The set S is identical to that of tables for which the sample odds ratio is at least as large as observed.

The hypergeometric applies directly as a sampling model when both sets of marginal counts are naturally fixed. A classic example of this case is Fisher's (1935b) tea-tasting experiment, relating to a woman's claim to be able to judge whether tea or milk was poured in the cup first. The woman was given eight cups of tea, in four of which tea was poured first and in four of which milk was poured first, and was told to guess which four had tea added first. The contingency table for this design (see Table 2) has $n = 8$ and $n_{1+} = n_{+1} = 4$; n_{11} can take values 0, 1, 2, 3, 4 with corresponding one-sided p-values (for H_1: $\theta > 1$) of 1.0, 0.986, 0.757, 0.243 and 0.014.

Ways of forming two-sided p-values in Fisher's exact test were discussed by Gibbons and Pratt (1975), Yates and discussants (1984), Davis (1986),

TABLE 2
Fisher's tea-tasting experiment

Poured first	Guess poured first: Milk	Guess poured first: Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Dupont (1986), Mantel (1987), and Lloyd (1988a). The most popular approaches are (a) double the one-sided p-value, (b) (2.1) with $S = \{t: f(t \mid n, n_{1+}, n_{+1}) \leq f(n_{11} \mid n, n_{1+}, n_{+1})\}$, and (c) (2.1) with $S = \{t: \mid t - E(n_{11})\mid \geq \mid n_{11} - E(n_{11})\mid\}$. The third option is identical to the null probability that the Pearson chi-squared statistic is at least as large as observed. Different approaches can give different results because of the discreteness and potential skewness of the hypergeometric distribution. For instance, consider the table having counts by row (10, 90/20, 80) (i.e., 10 and 90 in the first row, 20 and 80 in the second), discussed by Dupont (1986). The null distribution of n_{11} is symmetric about 15, and the three p-values are identical and equal 0.073. However, for the nearly identical table (10, 91/20, 80), p-value (a) equals 0.069, whereas (b) and (c) equal 0.050.

2.2 "Exact" Estimation in 2 × 2 Tables

The non-null conditional distribution (1.2) of n_{11} is used in constructing "exact" confidence intervals for the odds ratio. For data n_{11}, the conditional ML estimate of θ is the value of θ that maximizes probability (1.2). The estimate is obtained using iterative methods (Cornfield, 1956), and differs from the unconditional ML estimate $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$. For instance, for Fisher's tea-tasting data (Table 2), $n_{11} = 3$ and the unconditional ML estimate is 9; the conditional ML estimate maximizes the conditional likelihood (1.2),

$$16\theta^3/[1 + 16\theta + 36\theta^2 + 16\theta^3 + \theta^4],$$

and equals 6.41.

For testing H_0: $\theta = \theta_0$ against H_1: $\theta > \theta_0$, the p-value is $P = \sum_S f(t \mid n, n_{1+}, n_{+1}; \theta_0)$, where $S = \{t: t \geq n_{11}\}$. For testing against H_0: $\theta < \theta_0$, $S = \{t: t \leq n_{11}\}$. One can obtain an "exact" confidence interval for θ by inverting the test (Cornfield, 1956; Mantel and Hankey, 1971; Thomas, 1971). The lower endpoint is the θ_0 value for which $P = \alpha/2$ in testing H_0: $\theta = \theta_0$ against H_1: $\theta > \theta_0$. The upper endpoint is the θ_0 value for which $P = \alpha/2$ in testing H_0 against H_1: $\theta < \theta_0$. If $n_{11} = 0$, then the lower endpoint is 0, and one uses $P = \alpha$ in obtaining the upper endpoint; if $n_{11} = \min(n_{1+}, n_{+1})$, then the upper endpoint is ∞, and one uses $P = \alpha$ in obtaining the lower endpoint. For the tea-tasting data, the 95% confidence interval obtained in this manner is (0.21, 626.2).

The discreteness of the distribution of n_{11} limits the confidence intervals to a discrete set of possible endpoints, for fixed α. Thus, the true confidence coefficient is *at least* $1 - \alpha$, rather than *exactly* $1 - \alpha$, and its value depends on the value of θ (Neyman, 1935). It is strictly greater than $1 - \alpha$ unless the true θ is an attainable endpoint. I use quotes around "exact" in referring to such intervals to reflect this behavior.

Baptista and Pike (1977) described an alternative approach that also guarantees the confidence level but sometimes gives slightly shorter intervals. Their interval is a generalization of Sterne's (1954) confidence interval for a single binomial parameter. One inverts a family of acceptance regions that are formed using the minimal number of most likely outcomes. Specifically, for each θ_0, one finds a set $A(\theta_0)$ of possible n_{11} values such that the probability of the set is at least $1 - \alpha$ and such that every integer in the set is at least as likely to occur as every integer outside the set. The confidence interval is the set of θ_0 values for which the observed n_{11} falls in $A(\theta_0)$.

It is not possible to construct "exact" confidence intervals for association measures that are not functions of the odds ratio. They do not occur as parameters in generalized linear models with Poisson or binomial random component using canonical links. Thus, the usual conditioning arguments do not eliminate nuisance parameters. For instance, consider estimation of the difference of probabilities $\delta = \pi_1 - \pi_2$ for independent binomial samples. The joint sampling distribution can be expressed in terms of δ and π_1, for instance, but conditioning on the marginal totals does not eliminate π_1. Santner and Snell (1980) discussed this and other difficulties in interval estimation of the difference of proportions and the relative risk. They also described ways of getting conservative confidence intervals for these parameters, but the usual conservativeness due to discreteness is compounded because of the approach used to eliminate the nuisance parameter.

"Exact" interval estimates exist in certain extreme situations in which asymptotic interval estimates do not. Suppose the conditional inference uses the distribution of T_1, given T_0. When T_1 assumes its maximum or minimum value, the unconditional (or conditional) ML estimator of θ and its asymptotic standard error do not exist (unless one adds some constant to the cells), but "exact"

one-sided confidence intervals do exist. For a point estimate in such cases, some authors (e.g., Hirji, Mehta and Patel, 1987, 1988; Hirji, Tsiatis and Mehta, 1989) use median unbiased estimates. To illustrate, for a 2×2 table, when $n_{11} = 0$, the ML estimator of the loglinear association parameter (the log odds ratio) and the ML estimator $(\sum\sum 1/n_{ij})$ of the asymptotic variance of the ML estimator do not exist. The median unbiased estimate is the θ_0 value that gives $P = 0.5$ and the "exact" upper $100(1 - \alpha)\%$ confidence limit is the θ_0 value that gives $P = \alpha$, in testing against H_1: $\theta < \theta_0$.

2.3 Comparing Dependent Proportions

Next, consider the comparison of two proportions when each sample contains the same subjects, or the sample consists of matched pairs. For instance, one might have repeated measurement of a binary response at two occasions. Let n_{ij} be the number of observations in category i at occasion 1 and category j at occasion 2. Let π_{ij} denote the probability of response i at occasion 1 and response j at occasion 2, and let $p_{ij} = n_{ij}/n$. The sample proportions of "successes" p_{1+} at occasion 1 and p_{+1} at occasion 2 are dependent, rather than independent, because of the matching. The hypothesis of marginal homogeneity, H_0: $\pi_{1+} = \pi_{+1}$, corresponds to homogeneity for the 2×2 table consisting of the two marginal distributions. For binary responses, marginal homogeneity is equivalent to symmetry, $\pi_{12} = \pi_{21}$. To obtain an exact test, one conditions on $n^* = n_{12} + n_{21}$. Under H_0, n_{12} has a binomial $(n^*, 1/2)$ distribution. A two-sided p-value is the sum of binomial probabilities for n_{12} values at least as far from $n^*/2$ as observed. This is a small-sample version of *McNemar's test* (McNemar, 1947).

Cox (1958a) provided an argument, of which I now sketch the outline, that motivates this conditional approach. Let (Y_{1h}, Y_{2h}) denote the hth pair of observations, where a "1" response denotes category 1 (success) and "0" denotes category 2. Consider the logit model

$$\log[P(Y_{th} = 1)/P(Y_{th} = 0)] = \alpha_h + \beta I(t = 2) \tag{2.2}$$

where the indicator $I(t = 2)$ equals 1 when $t = 2$ and 0 when $t = 1$. This model (a special case of the *Rasch* model) permits separate response distributions for each pair, but assumes a common effect, $\exp(\beta)$ representing the odds ratio of success at occasion 2 compared with occasion 1. For the joint mass function of the data under the assumption that Y_{1h} and Y_{2h} are independent within each pair, one eliminates the nuisance parameters $\{\alpha_h\}$ by conditioning on the pairwise success totals $\{y_{1h} + y_{2h}\}$. Given a total of 0, $P(Y_{1h} = Y_{2h} = 0) = 1$, and given a total of 2, $P(Y_{1h} = Y_{2h} = 1) = 1$. The conditional distribution of (Y_{1h}, Y_{2h}) depends on β only when $y_{1h} + y_{2h} = 1$. For each of the n^* such pairs, direct calculation using (2.2) shows that $P(Y_{1h} = 1, Y_{2h} = 0) = [1 + \exp(\beta)]^{-1}$. Since $n_{12} = \sum y_{1h}$ for these n^* pairs, the distribution of n_{12} conditional on n^* is binomial with parameter $[1 + \exp(\beta)]^{-1}$. The parameter equals $1/2$ under marginal homogeneity ($\beta = 0$). This conditional analysis implies that pairs having identical response at the two occasions are irrelevant to testing $\pi_{1+} = \pi_{+1}$.

Cox (1966) generalized this analysis for data in which each case is matched with several controls. Gart (1969) gave an exact test for comparing matched proportions in crossover designs.

3. EXACT INFERENCE IN $I \times J$ TABLES

For $I \times J$ tables, statistical independence of X and Y corresponds to the loglinear model $M_0 = (X, Y)$, which has form

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y.$$

To test this model against the saturated model [M_1 = (1.1) extended to I rows and J columns], one uses the null distribution of $\{n_{ij}\}$ given $\{n_{i+}\}$ and $\{n_{+j}\}$.

For multinomial sampling, the cell probability parameters $\{\pi_{ij}\}$ in the distribution of $\{n_{ij}\}$ can be expressed in terms of the marginal probabilities and $(I - 1)(J - 1)$ odds ratios. Conditional on $\{n_{i+}\}$ and $\{n_{+j}\}$, Cornfield (1956) noted that the distribution of $\{n_{ij}\}$ depends only on the odds ratios. The conditional probabilities are proportional to

$$\frac{\prod_{i=1}^{I-1}\prod_{j=1}^{J-1}\alpha_{ij}^{n_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!}$$

where $\alpha_{ij} = (\pi_{ij}\pi_{IJ})/(\pi_{iJ}\pi_{Ij})$. For exact tests of statistical independence (all $\alpha_{ij} = 1$), this distribution simplifies to the multiple hypergeometric. The probability of a table $\{n_{ij}\}$ having the given marginal totals equals

$$\frac{\left(\prod_i n_{i+}!\right)\left(\prod_j n_{+j}!\right)}{n!\prod_i\prod_j n_{ij}!}. \tag{3.1}$$

3.1 Linking Test Statistics to Alternatives

Let p_{obs} denote the null probability (3.1) of the observed table. Freeman and Halton (1951) defined the p-value for a conditional test of independence to be the null probability of the set of tables having probability no greater than p_{obs}. For the 2×2 case, this simplifies to a two-sided version of Fisher's exact test. Many statisticians have argued that the p-value should instead be based on the exact distribution of some meaningful statistic T (such as the efficient score statistic) for quantifying the departure of the data from H_0 (e.g., Yates, 1934; Fisher, 1950; Healy, 1969; Agresti and Wackerly, 1977; Cressie and Read, 1989). The p-value for the Freeman-Halton test may be regarded as one that uses a statistic negatively related to p_{obs}, such as $T = -2\log(p_{obs})$.

When both classifications are nominal, the usual alternative to the independence model is the saturated model. The efficient score statistic is then the Pearson chi-squared statistic for testing the goodness-of-fit of the independence model,

$$X^2 = \sum_i \sum_j \left[n_{ij} - \hat{m}_{ij}\right]^2 / \hat{m}_{ij},$$

where $\hat{m}_{ij} = n_{i+}n_{+j}/n$. One could let the p-value be the null probability that X^2 is at least as large as observed (Yates, 1934; Agresti and Wackerly, 1977; Baker, 1977). More generally, one could use a power divergence statistic (Cressie and Read, 1989), special cases of which are the Pearson statistic and the likelihood-ratio statistic.

When both classifications are ordinal, it is often important to have good power for detecting a monotone trend in the association. To do this, one could let T relate to a Spearman or Pearson-type correlation (Patefield, 1982), the difference between the numbers of concordant and discordant pairs (Agresti and Wackerly, 1977), or the Jonckheere-Terpstra statistic (StatXact, 1991). In the first case, the statistic on which the reference tables are ordered is $T = \sum\sum x_i y_j (n_{ij} - n_{i+}n_{+j}/n)$, for two sets of monotone scores $\{x_i\}$ for rows and $\{y_j\}$ for columns. An exact conditional test using this statistic is an efficient score test for the loglinear model of linear-by-linear association (Agresti, Mehta and Patel, 1990). That model has the form

$$(3.2) \qquad \log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta x_i y_j,$$

(Birch, 1965; Haberman, 1974; Goodman, 1979), with the special case $\beta = 0$ representing statistical independence of X and Y. For model (3.2), the sufficient statistics are T plus those for the independence model $(\{n_{i+}\}, \{n_{+j}\})$. Cohen and Sackrowitz (1991) showed that nonrandomized tests using T fall in a complete class of unbiased and admissible tests.

When X is nominal and Y is ordinal, one might let T be the Kruskal-Wallis statistic adjusted for ties (Klotz and Teng, 1977). In that statistic, average ranks are used as scores $\{y_j\}$ for the column categories, and the test is sensitive to variation among the mean ranks computed for the conditional distributions within the rows. This is an efficient score statistic for a loglinear model having form (3.2), with "row effects" $\{\beta x_i\}$ treated as parameters. For $2 \times J$ tables, Soms (1985) presented exact tests sensitive to other alternatives. See Haberman (1974), Goodman (1979) and Agresti (1990, Chapter 8) for discussions of loglinear models with scores.

To illustrate that tables that are "more contradictory" to H_0 according to some statistic T need not be less likely, consider the twelve 3×3 tables having row totals (6, 1, 2) and column totals (1, 2, 6). Table 3 gives the conditional null distribution of $T = \sum\sum x_i y_j n_{ij}$, for $\{x_i = y_i = i - 2\}$. The distribution has support between -7 and 0 with a mean of -2.22. It is highly irregular, far from its limiting normal distribution. Table 3 shows that, for some margins, it is not possible to obtain small p-values for some alternatives (e.g., a test using T for the one-sided alternative of a positive association). The table having entries by row (0, 2, 4/1, 0, 0/0, 0, 2), one of two tables to have $T = -2$, has only a quarter of the probability of the table having entries (1, 2, 3/0, 0, 1/0, 0, 2), the only table having $T = 0$. Yet, the first table is less contradictory to H_0 than the second, using $|T - E(T)|$ as the criterion. The first table has a p-value of 1.0 using this criterion, but only 0.2856 in the Freeman-Halton test. By contrast, the $I \times I$ table having $n_{ii} = 1$ for all i and $n_{ij} = 0$ for all $i \neq j$ has $P = (2/I!)$ using $|T - E(T)|$ but has $P = 1.0$ for the Freeman-Halton test or the exact test using X^2 as the criterion.

TABLE 3
Exact conditional distribution of $T = \sum\sum(i-2)(j-2)n_{ij}$ *under independence, for margins* (6, 1, 2) *and* (1, 2, 6)

t	No. of tables	Probabilities	$P(T = t)$
0	1	0.2381	0.2381
−1	1	0.1190	0.1190
−2	2	0.0595, 0.2381	0.2976
−3	1	0.0476	0.0476
−4	3	0.0238, 0.0476, 0.1190	0.1904
−5	1	0.0476	0.0476
−6	2	0.0040, 0.0476	0.0516
−7	1	0.0079	0.0079

When $I = 2$, ordering the tables by T is equivalent to ordering them by $U = \sum_j y_j n_{1j}$. Many statistical tests use U for various choices of scores (Graubard and Korn, 1987). For arbitrary monotone scores, the exact test for T is a small-sample version of a trend test proposed by Cochran (1954) and Armitage (1955). The statistic with midrank scores is used in exact Wilcoxon tests for ordered categorical data (Klotz, 1966; Mehta, Patel and Tsiatis, 1984).

3.2 Exact Estimation for $I \times J$ Tables

"Exact" confidence intervals are rarely used for $I \times J$ tables, probably because of the complexity. Reducing parameter dimensionality could be useful in many applications, by using an unsaturated loglinear model. Agresti, Mehta and Patel (1990) discussed confidence intervals for ordinal classifications when odds ratios satisfy the pattern

$$\log \alpha_{ij} = \beta(x_I - x_i)(y_J - y_j)$$

implied by the linear-by-linear association model (3.2). For this pattern, the non-null distribution of $T = \sum\sum x_i y_j n_{ij}$ is

$$P(T = t \mid \{n_{i+}\}, \{n_{+j}\}; \beta) = \frac{C_t e^{\beta t}}{\sum_u C_u e^{\beta u}}$$

where C_t is the sum of $(\Pi_i \Pi_j n_{ij}!)^{-1}$ for all tables with the given marginal distributions having $T = t$. One obtains a confidence interval for β and thus $\{\alpha_{ij}\}$ by inverting the test of H_0: $\beta = \beta_0$, as in the case of the odds ratio for a 2×2 table. When T takes its maximum or minimum possible value for the given marginals, the conditional and unconditional ML estimators of β and asymptotic standard errors do not exist, but one-sided confidence bounds for β and odds ratios using it do exist.

4. EXACT INFERENCE IN THREE-WAY CONTINGENCY TABLES

I next consider three-way tables, paying special attention to the $2 \times 2 \times K$ case. Such tables occur, for instance, when one compares a binary response (Y) for two treatments (X), using data obtained at K levels of a possibly confounding factor (Z). Two hypotheses of importance are (1) conditional independence of X and Y, given Z, and (2) no three-factor interaction, meaning that the true X-Y odds ratio is identical at each level of Z. Conditional independence of X and Y, given Z, corresponds to the loglinear model symbolized by (XZ, YZ), and no three-factor interaction corresponds to the loglinear model (XY, XZ, YZ). For cell-expected frequencies $\{m_{ijk}\}$, model (XY, XZ, YZ) has the form

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

and model (XZ, YZ) is its special case in which $\{\lambda_{ij}^{XY} = 0\}$.

4.1 Testing Conditional Independence in $2 \times 2 \times K$ Tables

A common approach to testing conditional independence [$M_0 = (XZ, YZ)$] performs the analysis under the assumption of no three-factor interaction [$M_1 = (XY, XZ, YZ)$]. The sufficient statistics are the X-Z and Y-Z two-way marginal tables for M_0, and these as well as the X-Y marginal table for M_1. The relevant conditional distribution is that of the X-Y marginal table, given the X-Z and Y-Z marginal tables. The conditioned totals are the marginal counts for the K partial tables. For the $2 \times 2 \times K$ case, the distribution simplifies to that of $T = \sum_k n_{11k}$, given $\{n_{1+k}, n_{2+k}, n_{+1k}, n_{+2k}, k = 1, \ldots, K\}$. Under the assumption of no three-factor interaction, Birch (1964) showed that UMPU tests of conditional independence utilize T (see also Lehmann 1986, pages 163-164).

Conditional on the strata margins, $\{n_{11k}, k = 1, \ldots, K\}$ have independent hypergeometric distributions, each of form (1.2) with $\theta = 1$. The product of the K mass functions determines the null distribution of their sum. For the one-sided alternative of a "positive" association (odds ratio greater than 1.0 in each level of Z), the p-value for Birch's exact test of conditional independence is the null probability that $\sum_k n_{11k}$ is at least as large as observed, for the fixed marginal totals. Thomas (1975), Pagano and Tritchler (1983b), and Mehta, Patel and Gray (1985) gave algorithms for implementing this test, which may be regarded as an exact small-sample version of the Cochran-Mantel-Haenszel test.

The exact McNemar test for matched pairs (Section 2.3) is a special case of Birch's exact test of conditional independence. In this representation, each of the n matched pairs has a 2×2 table relating occasion (or member of pair) to response. One tests M_0 = conditional independence of occasion and response, given the pair, under the assumption of M_1 = homogeneous odds ratios in the n 2×2 tables.

For $2 \times 2 \times K$ tables in which it is unrealistic to expect the K conditional X-Y odds ratios to be similar or even of the same sign, the saturated model is a more relevant alternative than no three-factor interaction. In that case, the efficient score statistic is $\sum_k X_k^2$, where X_k^2 denotes the Pearson statistic for testing independence between

X and Y within the kth level of Z. This statistic is commonly used in asymptotic tests of conditional independence against the general alternative. Currently, there do not seem to be any computer algorithms for this case.

4.2 Testing Homogeneity of Odds Ratios in $2 \times 2 \times K$ Tables

Birch's exact test using test statistic $\sum_k n_{11k}$ assumes homogeneity of the odds ratios in the $2 \times 2 \times K$ table. Zelen (1971) presented an exact test of this assumption. Here, $M_0 = (XY, XZ, YZ)$, and M_1 is the saturated model. The conditional distribution, given each of the three sets of two-way marginal totals, has probabilities proportional to $(\Pi_i \Pi_j \Pi_k n_{ijk}!)^{-1}$. For the $2 \times 2 \times K$ case, the fixed totals are $\{n_{1+k}, n_{2+k}, n_{+1k}, n_{+2k},\ k = 1, \ldots, K\}$ and $\{n_{11+}, n_{12+}, n_{21+}, n_{22+}\}$. Zelen defined the p-value to be the sum of probabilities of all $2 \times 2 \times K$ tables that are no more probable than the observed table. Alternatively, one could define P by ordering the tables with the given two-way margins by X^2 for testing the fit of loglinear model (XY, XZ, YZ). Thomas (1975) and Pagano and Tritcher (1983b) gave algorithms for implementing Zelen's test.

To improve potential power in testing model (XY, XZ, YZ), one could instead test it against an unsaturated model. When the levels of Z have a natural ordering, one could use the alternative loglinear model by which the log odds ratios change linearly across the K strata; that is,

$$\log m_{ijk} = (XY, XZ, YZ) + I(i = j = 1)\, \delta z_k,$$

for fixed monotone scores $\{z_k\}$, where $I(\cdot)$ is the indicator function. The relevant distribution is then that of $\sum_k z_k n_{11k}$, conditional on all two-way marginal totals. Zelen (1971) also presented such an exact test.

To illustrate some exact tests for conditional independence and for homogeneity of odds ratios for $2 \times 2 \times K$ tables, consider Table 4, a $2 \times 2 \times 3$ table based on a larger table presented by Gastwirth (1988, page 266). The data refers to P = whether promoted and R = race, stratified by M = month of promotion consideration (in 1974). Cases involved GS-13 level computer specialists being considered for promotion to level GS-14. (It appears that many of the subjects appeared in two or all three strata, but this information is not available. So, like Gastwirth, I treat the promotion decisions as independent.) Under the assumption of a constant odds ratio θ between P and R at each level of M, we first test H_0: $\theta = 1$ against H_1: $\theta < 1$. In using the one-sided alternative for the association between P and R, the test is sensitive to evidence of possible discrimination against blacks, in the sense of the probability of promotions being lower for blacks than whites. Given the P and R marginal totals at each level of M, n_{111} can range between 0 and 4, n_{112} can range between 0 and 4, and n_{113} can range between 0 and 2. The test statistic $T = \sum n_{11k}$ can assume values between 0 and 10, and under H_0, $E(T) = 2.90$ and $\sigma(T) = 1.35$. Note that the sample data represent the most extreme possible result in each of the three cases. The observed test statistic is $T = 0$, and the p-value is the null value of $P(T \leq 0)$, which is 0.026. Zelen's test of the hypothesis of a constant odds ratio is a test of fit of the loglinear model (PR, PM, RM). The reference set consists of the subset of these $2 \times 2 \times 3$ tables that also satisfy $n_{11+} = 0$. The observed table is the only such table, so the test is degenerate, giving $P = 1.0$.

TABLE 4
Example of $2 \times 2 \times K$ *analysis*

	July promotions		August promotions		September promotions	
Race	Yes	No	Yes	No	Yes	No
Black	0	7	0	7	0	8
White	4	16	4	13	2	13

Source: Gastwirth (1988).

4.3 Exact Estimation in $2 \times 2 \times K$ Tables

Assuming no three-factor interaction and conditioning on the strata totals, the joint distribution of $\{n_{111}, \ldots, n_{11K}\}$ is the product of K terms of the type given in (1.2) for the 2×2 case. This distribution depends only on the common odds ratio. Birch (1964) discussed the conditional ML estimator of the common odds ratio, and Gart (1970) defined "exact" confidence intervals as a direct extension of Cornfield's "exact" intervals for 2×2 tables. Thomas (1975), Pagano and Tritchler (1983b), Mehta, Patel and Gray (1985), and Vollset, Hirji and Elashoff (1991) provided algorithms for calculating such intervals. When the sufficient statistic $T = \sum_k n_{11k}$ attains its minimum or maximum possible value, asymptotic confidence intervals (e.g., using the Mantel-Haenszel approach; see Agresti 1990, pages 235-236) do not exist, but one-sided "exact" confidence intervals are well defined.

Table 4 has the boundary value $T = 0$, resulting in a degenerate conditional ML estimate of 0.0 for an assumed constant odds ratio. An "exact" 95% upper confidence bound for the odds ratio equals 0.779, the value θ_0 that gives $P = 0.05$ in testing

H_0: $\theta = \theta_0$ against H_1: $\theta < \theta_0$. The median unbiased estimator equals 0.152.

4.4 Exact Methods for $I \times J \times K$ Tables

In principle, methods for exact testing and estimation extend to loglinear models for multiway tables. Current computational algorithms are restricted to certain analyses for $2 \times J \times K$ tables. In this subsection, I outline the types of inferences for which it would be useful to extend computational algorithms in the $I \times J \times K$ case.

Consider the hypotheses corresponding to the following five hierarchical loglinear models:

(1) (X, Y, Z): Mutual independence of X, Y, and Z;
(2) (X, YZ): Joint independence of X and the Y-Z classification;
(3) (XZ, YZ): Conditional independence of X and Y, given Z;
(4) (XY, XZ, YZ): No three-factor interaction;
(5) (XYZ): Saturated model.

For each of models (1) to (4), one can consider exact testing against the alternative of the next most complex model, and against the general alternative of the saturated model.

Certain tests are special cases of ones already developed for two-way tables, so they do not require separate consideration. Hypothesis [2: (X, YZ)] is a special case of statistical independence for a two-way table, in which the second classification consists of the JK combinations of categories of Y and Z. Thus, an exact test of [2: (X, YZ)] against [5: (XYZ)] is simply a standard exact test of independence for a two-way ($I \times JK$) table. For instance, in Table 4, an exact test of (P, RM) (i.e., that promotion is jointly independent of race and month of decision) is an exact test for the two-way table having rows $(0, 4, 0, 4, 0, 2/7, 16, 7, 13, 8, 13)$. The Pearson test statistic of 5.62 has an exact p-value of 0.353. [The attained significance for this joint test is weak compared to the p-value of 0.026 obtained for the P-R association alone in the one-sided exact test of (PM, RM) against (PR, PM, RM). This shows how severely the evidence represented by an effect of a certain size can diminish when the degrees of freedom on which it is based increase drastically.]

A test of [2: (X, YZ)] against [3: (XZ, YZ)] tests whether the λ^{XZ} term in model (XZ, YZ) is zero. By standard collapsibility results (e.g., Bishop, 1971; Agresti, 1990, pages 146 and 230), when this model holds the λ^{XZ} term is identical to the λ^{XZ} term in the saturated two-way loglinear model for the marginal X-Z two-way table. Thus, one can conduct an exact test of [2: (X, YZ)] against [3: (XZ, YZ)] using an exact test of independence of X and Z in that two-way table. For instance, one can conduct an exact test of (P, RM) against (PR, RM) by simply testing whether P and R are independent in the marginal table $(0, 22/10, 42)$; for this table, the Pearson statistic has two-sided $P = 0.056$. Similarly, one can conduct an exact test of [1: (X, Y, Z)] against [2: (X, YZ)] using an exact test of independence of Y and Z in the two-way Y-Z marginal table. Computational algorithms are unavailable for the other situations, discussed in the remainder of this subsection.

To test hypothesis [1: (X, Y, Z)] against the saturated model, one conditions on the sufficient statistics for (X, Y, Z), which are $\{n_{i++}, n_{+j+}, n_{++k}\}$. The resulting mass function is

$$\frac{\left(\prod_i n_{i++}!\right)\left(\prod_j n_{+j+}!\right)\left(\prod_k n_{++k}!\right)}{(n!)^2 \prod_i \prod_j \prod_k n_{ijk}!}.$$

Stumpf and Steyn (1986) gave formulas for the first- and second-order moments of the cell counts. To construct an exact test of hypothesis [1: (X, Y, Z)] against the saturated model, one could use this distribution to generate the exact conditional distribution of the Pearson statistic for testing the fit of model (X, Y, Z). This case is relatively unimportant, as the hypothesis of mutual independence is plausible in very few applications.

A more important case is a test of conditional independence [3: (XZ, YZ)] of X and Y against [4: (XY, XZ, YZ)]. Such a test generalizes Birch's test for $2 \times 2 \times K$ tables. Here, one tests conditional independence under the assumption that the $(I - 1)(J - 1)$ odds ratios relating X and Y are identical across the K levels of Z. Model (XZ, YZ) has the sufficient statistics $(\{n_{i+k}\}, \{n_{+jk}\})$, and model (XY, XZ, YZ) has these plus $\{n_{ij+}\}$. One uses the distribution of $(\{n_{ij+}\} \mid \{n_{i+k}\}, \{n_{+jk}\})$, which is a product of multivariate hypergeometric distributions from the various layers of Z. Let $\mathbf{d}$ denote the $(I - 1)(J - 1) \times 1$ vector having elements

$$d_{ij} = \sum_k \left[n_{ijk} - \left(\frac{n_{i+k} n_{+jk}}{n_{++k}} \right) \right], \quad i = 1, \ldots, I - 1, j = 1, \ldots, J - 1.$$

The efficient score test orders tables in the reference set using $\mathbf{d}'\mathbf{V}^{-1}\mathbf{d}$, where $\mathbf{V}$ is the null covariance matrix of $\mathbf{d}$. Birch (1965) proposed an asymptotic test of this type.

Alternatively, one could test conditional independence [3: (XZ, YZ)] against [5: (XYZ)], if one believes that the association between X and Y may vary considerably across levels of Z. An efficient score statistic is then the Pearson statistic for testing the fit of (XZ, YZ), which is $\sum_k X_k^2$, where X_k^2 is the Pearson statistic for testing independence between X and Y within the kth level of Z. This test would also use the conditional distribution that fixes the XZ and YZ marginal tables. Since it has a broader alternative than the generalized Birch test, this test would tend to be less powerful than that test when model (XY, XZ, YZ) provides a decent approximation to the actual distribution.

An exact test of no three-factor interaction [4: (XY, XZ, YZ)] for $I \times J \times K$ tables generalizes Zelen's test for $2 \times 2 \times K$ tables. Here, the null hypothesis states that the $(I-1)(J-1)$ odds ratios relating X and Y are identical across the K levels of Z. The relevant conditional distribution has probabilities proportional to $(\Pi_i \Pi_j \Pi_k n_{ijk}!)^{-1}$, for the reference set of tables having XY, XZ and YZ marginal tables identical to the observed ones. An efficient score test against the general alternative [5: (XYZ)] orders tables in the reference set by the Pearson statistic for testing the fit of the model.

For ordinal variables, one would modify the above ideas by constructing tests to increase power against important alternatives, such as has been done for $I \times J$ tables. For instance, to test conditional independence (XZ, YZ), one might choose an alternative model that implies a monotone conditional X-Y association. One could use the single degree-of-freedom test statistic $\sum_k \{\sum_i \sum_j x_i y_j [n_{ijk} - (n_{i+k} n_{+jk})/n_{++k}]\}$ to order the reference set of tables having the fixed X-Z and Y-Z marginal tables. This results from comparing model (XZ, YZ) to a model of homogeneous linear-by-linear association, whereby the λ_{ij}^{XY} term in model (XY, XZ, YZ) is replaced by a term $\beta x_i y_j$ for fixed monotone row and column scores. Such a test would be an exact analog of an asymptotic score test proposed by Mantel (1963) and Birch (1965).

There has been some work of this type for the $2 \times J \times K$ case with ordered levels of Y (Mehta, Patel and Senchaudhuri, 1991). In this case, one can take $x_1 = 1$ and $x_2 = 0$, and consider the conditional distribution of $\sum_k [\sum_j y_j n_{ijk}]$. For preselected monotone scores, this gives a stratified version of the Cochran–Armitage trend test. For rank $\{y_j\}$ scores, it gives a stratified version of the Wilcoxon test. Another application of this case is testing marginal homogeneity in an $I \times I$ table with the same ordered row and column categories. One can conduct this test by conducting the exact test of conditional independence for the $2 \times I \times n$ table, where the two rows for stratum k contain one observation in each row, giving the responses at the two occasions for subject k. Such a test is an exact analog of asymptotic tests described by White, Landis and Cooper (1982) and Kuritz, Landis and Koch (1988).

Interval estimation for $I \times J \times K$ tables would seem to be awkward except for simpler models that reduce the dimensionality of the parameter space. Examples include models that describe the X-Y conditional association by a linear-by-linear term or describe three-factor interaction by a linear trend in conditional log odds ratios.

5. EXACT INFERENCE IN LOGISTIC REGRESSION MODELS

Exact inferences for loglinear models discussed in previous sections have counterparts for other generalized linear models that use natural parameters as the basis of the link function. A closely related example for categorical data is logistic regression modeling. Assuming a binomial distribution with parameter π for the response, one uses the logit link, $\log[\pi/(1-\pi)]$. Logistic models are particularly useful when highlighting one categorical variable in the contingency table as a response and the others as explanatory. Loglinear models do not make this distinction, although logistic models with qualitative explanatory variables have equivalent loglinear representations.

For subject i, let y_i denote a binary response, and let $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$ denote values of k explanatory variables, where $x_{i0} = 1$. The logistic regression model is

$$(5.1) \qquad \pi(\mathbf{x}_i) = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}$$

or

$$\log[\pi(\mathbf{x}_i)/(1 - \pi(\mathbf{x}_i))] = \sum_j \beta_j x_{ij}.$$

Under the usual assumption that $\{y_i\}$ are independent Bernoulli outcomes, the sufficient statistic for β_j is $T_j = \sum_i y_i x_{ij}$, $j = 0, \ldots, k$. As noted by Cox (1958b, 1970), one can conduct exact inference for β_j using the distribution of T_j, conditional on $\{T_i, i \neq j\}$. Such inference is called *conditional logistic regression* (Bayer and Cox, 1979; Breslow and Day, 1980, Chapter 7; Tritchler, 1984; Hirji, Mehta and Patel, 1987).

To illustrate, Table 5 shows some data from a case-control study (Shapiro et al., 1979) relating cigarette smoking to myocardial infarction for women of various ages using oral contraceptives.

TABLE 5
Example for exact logistic analysis

Age	Disease status	Smoking level (cigarettes/day) 0	1-24	> 24
25-29	Myocardial infarction	0	1	3
	Control	25	25	12
30-34	Myocardial infarction	0	1	8
	Control	13	10	10

Source: Shapiro et al. (1979).

Let $\{n_{ijk}\}$ denote the count at level i of smoking, j of disease status, and k of age. Let $\{x_i\}$ denote scores assigned to the levels of smoking. Let π_{ik} denote the probability of disease for subjects at level i of smoking and k of age. One might consider the model

$$(5.2)\quad \log\left[\frac{\pi_{ik}}{1-\pi_{ik}}\right] = \beta_0 + \beta_1 x_i + \beta_2 I(\text{age} = 2).$$

Then $\{n_{i+k}\}$ are fixed for this model, and $\{n_{+jk}\}$ are fixed by the retrospective nature of the study. To conduct exact inference about β_1, one considers the distribution of $\sum_k(\sum_i x_i n_{i1k})$ (the sufficient statistic for β_1), conditional on these totals. For $\{x_1 = 0,\ x_2 = 12.5,\ x_3 = 30\}$, the conditional ML estimate of β_1 is 0.130, and the exact p-value for testing $\beta_1 = 0$ against $\beta_1 > 0$ is 0.000. Here, results are similar to those obtained with the unconditional ML analysis, for which the estimate of β_1 of 0.133 has an estimated standard error of 0.042. One could add an interaction term to the model and do an exact analysis for it. This would be particularly natural with additional age strata representing greater variation in the age factor, since one might expect the effect of smoking to increase at higher age levels.

Several special cases of exact analyses for the logistic model have received attention in the literature. For instance, Breslow and Day (1980), Peritz (1982), and Hirji, Mehta and Patel (1988) discussed exact inference for matched case-control studies with the logistic model, in which case a subset of the explanatory variables are used for matching. When $k = 1$ in (5.1), the exact test of $\beta_1 = 0$ using $T_1 = \sum_i y_i x_{i1}$ given $T_0 = \sum_i y_i$ is a special case of the linear-by-linear test described in Section 3.1 applied to $I \times 2$ tables. Here, I represents the number of distinct sample values of the explanatory variable, and the test may be regarded as an exact version of the Cochran–Armitage trend test.

Difficulties can arise in exact inference for logistic regression when some explanatory variables are continuous. The $\{y_i\}$ values may be completely determined by the given sufficient statistics, making the conditional distribution degenerate.

Algorithms for exact conditional inference for logistic regression can be applied to perform inference for equivalent loglinear models. To illustrate, consider loglinear modeling of several $I \times 2$ tables. Regard the tables as a three-way $I \times 2 \times K$ cross-classification of X, Y and Z. The loglinear model is equivalent to a logistic model for response Y whenever that loglinear model has a general association term relating X and Z. For instance, the logistic model (5.2) for Table 5 corresponds to the loglinear model having form

$$\log m_{ijk} = \mu + \lambda_i^S + \lambda_j^D + \lambda_k^A + \lambda_{ik}^{SA} + \lambda_{jk}^{DA} + \beta_1 x_i y_j$$

where $\{y_1 = 1,\ y_2 = 0\}$ and S = smoking, D = disease status, and A = age. This is a special case of the loglinear model of homogeneous linear-by-linear S-D association.

6. EXACT GOODNESS OF FIT

One can interpret tests of independence, conditional independence and no three-factor interaction against *general* alternatives as tests of goodness of fit of loglinear models. In principle, one could use analogous methods to construct exact tests of goodness of fit for other loglinear or logistic regression models. For a particular model M, the reference set consists of all tables having the observed values for the minimal sufficient statistics. Given that the model holds, the conditional distribution of the data given those sufficient statistics is independent of any parameters. One could construct the test by computing the null distribution of a goodness-of-fit statistic, such as the Pearson statistic. The p-value for testing the model is the conditional probability that the goodness-of-fit statistic is at least as large as observed. The ML fitted values for the model are the same for all tables in the conditional reference set.

For instance, in testing independence with Fisher's exact test, one also implicitly tests the adequacy of the loglinear model of independence, (X, Y). To test the more general model (3.2) of linear-by-linear association in a two-way table, one considers the conditional distribution of the data given $\{n_{i+}\}$, $\{n_{+j}\}$, and $\sum\sum x_i y_j n_{ij}$.

McCullagh (1986) showed that, even for large samples, it is beneficial to perform goodness-of-fit tests using the conditional rather than unconditional distribution. Although an exact goodness-of-fit test makes theoretical sense for any model having simple sufficient statistics, a general

computer algorithm is not available for it. This is a useful topic for future research. There is also a need for work on exact distributions of localized measures of goodness of fit, such as cell residuals. Bedrick and Hill (1990) gave exact conditional tests for a single outlier and for multiple outliers in logistic regression.

7. COMPUTING FEASIBILITY

Doing computations for exact conditional inference requires working with the set of contingency tables having the given values of the sufficient statistics that are fixed for the inference. The potentially huge cardinality of the conditional reference set has been a severe impediment to the use of exact tests.

To illustrate, a 4×4 table with only 20 observations can have as many as 40,176 tables with the same margins; a 4×4 table with 100 observations has a maximum cardinality on the order of 7.2×10^9. For given I and J and marginal proportions, the number of tables in the reference set of $I \times J$ tables with those fixed proportions increases exponentially in the sample size n. For fixed n, the number of tables having given row and column marginal proportions also increases rapidly as I and J increase or as the row and column proportions become more homogeneous. For instance, a 5×5 table has a maximum cardinality on the order of 2.1×10^6 for 20 observations and 9.2×10^{14} for 100 observations. Good (1976, 1977) and Gail and Mantel (1977) gave approximations for the cardinality, and Agresti and Wackerly (1977) and Agresti, Wackerly and Boyett (1979) gave maxima for several table dimensions and sample sizes.

Enormous improvements achieved recently both in algorithms and in computer power have made exact inference much more feasible than it was a decade ago. Most analyses conducted then with a mainframe computer can be conducted now in the same order of time on a personal computer. When one is interested only in a p-value rather than the entire distribution of some statistic, substantial savings in time are obtained using algorithms that do not require total enumeration of the reference set (Pagano and Halvorsen, 1981; Mehta and Patel, 1983). With some algorithms, exact analyses can be easily conducted on a PC running on MS-DOS when the order of the cardinality is about 10^7. They can be conducted when the cardinality is much larger, using computers having operating systems with larger memory capacities.

To illustrate, consider the 3×4 table $(60, 4, 1, 0/1, 5, 4, 1/3, 3, 3, 2)$, having $n = 100$ and 33,675 tables in the reference set. In 1978, I performed the Freeman-Halton test using a state-of-the-art FORTRAN program on an IBM 370/165 in about 15 seconds CPU time. This year, I performed the same analysis in 15 seconds total time using the software package StatXact (1991) on a 386-version PC (the CompuAdd 386SX, with math coprocessing chip) running on MS-DOS at 16 mHz, and in 1 second of CPU time using SAS (PROC FREQ) on a workstation (DEC 3100). The 4×4 table $(7, 5, 0, 0/1, 15, 1, 0/0, 7, 7, 0/0, 0, 4, 9)$ has 12,798,781 tables in the reference set. In 1977, Klotz and Teng (1977) estimated it would take 6 hours on a Univac 1110 at the University of Wisconsin to perform an exact Kruskal–Wallis test for a table having these margins. I performed this analysis with StatXact in about 8 minutes of total time on a 386 version PC. This table has a very small p-value (0.0000 rounded to four decimal places), which results in considerable savings in time for algorithms that are able to determine whether many tables contribute to the p-value without explicitly enumerating all their cells.

7.1 Algorithms

A variety of algorithms have been used in computing exact conditional distributions. Verbeek and Kroonenberg (1985) presented a good survey of those used for $I \times J$ contingency tables. These include algorithms that provide total enumeration of the tables in the reference set (e.g., March, 1972; Boulton, 1974; Baker, 1977; Cantor, 1979; Balmer, 1988), algorithms that compute the characteristic function and invert it via Fourier transforms (e.g., Pagano and Tritchler, 1983a, b), network algorithms (Mehta and Patel, 1983) and Monte Carlo algorithms (e.g., Agresti, Wackerly, and Boyett, 1979). Their paper also has enlightening discussions of practical problems related to the algorithms and hardware for implementing them, such as how to ensure proper comparison of extremely small probabilities.

Algorithms that provide total enumeration of the reference set are very time-consuming, and adequate only for small problems. In the characteristic function approach (Good, Gover and Mitchell, 1970; Good, 1982; Pagano and Tritchler, 1983a, b), one computes the characteristic function of the statistic of interest (such as a goodness-of-fit statistic) using a recurrence relation and then inverts it using a Fourier transform to obtain the relevant distribution. The fast Fourier transform (Cooley and Tukey, 1965) is a popular method for fast convolution of long sequences, and so it is a natural one to apply to analyses (such as those for $2 \times 2 \times K$ tables) involving convolutions of distributions. This method is relatively space and time efficient, computation

time increasing polynomially in the sample size, rather than exponentially. However, Vollset, Hirji and Elashoff (1991) noted that this method's use of trigonometric functions and complex arithmetic can introduce substantial round-off errors for non-null calculations when there is a wide range between the largest and smallest of the combinatorial coefficients that determine the exact distribution (e.g., a ratio of the two exceeding about 10^{20}).

Among the most popular and versatile programs developed in the past decade have been ones using the network algorithm. This algorithm has been applied to several problems in a series of papers by Cyrus Mehta, Nitin Patel and some coworkers. For instance, see Mehta and Patel (1983, 1986) for its application to Freeman–Halton exact tests for $I \times J$ tables; Mehta, Patel and Tsiatis (1984) for exact tests for $2 \times J$ tables; Mehta, Patel and Gray (1985) for inference for the common odds ratio in $2 \times 2 \times K$ tables; Hirji, Mehta and Patel (1987) for exact logistic regression; Agresti, Mehta and Patel (1990) for exact inference in $I \times J$ tables with ordered categories; Mehta, Patel and Senchaudhuri (1991) for inference for $2 \times J \times K$ tables; and Hilton, Mehta and Patel (1991) for Smirnov tests for categorical or continuous data.

I provide here only a brief outline of the network representation for the reference set for an $I \times J$ table with fixed margins, and refer the reader to the previously mentioned papers for technical details on the algorithm itself. The network representation consists of *nodes* and *arcs*, constructed in $J + 1$ stages. For $k = 0, \ldots, J$, the nodes at stage k have the form $(k, \mathbf{w}_k)$, where $\mathbf{w}_k = (w_{1k}, \ldots, w_{Ik})$, with $w_{ik} = n_{i1} + \cdots + n_{ik}$ and $\mathbf{w}_0 = \mathbf{0}$. There are as many nodes at stage k as there are possible partial sums for the first k columns of the table. Arcs emanate from each node at any stage k, each arc being connected to a distinct node at stage $k + 1$. The network is constructed recursively by specifying all successor nodes $(k + 1, \mathbf{w}_{k+1})$ that are connected by arcs to each node $(k, \mathbf{w}_k)$. A path through the network is a sequence of arcs $(0, \mathbf{0}) \rightarrow (1, \mathbf{w}_1) \rightarrow \cdots \rightarrow (J, \mathbf{w}_J)$. Each path represents a distinct table in the reference set, with entries $(\mathbf{w}_{k+1} - \mathbf{w}_k)$ in column $k + 1$. The network representation is used in calculating the exact distribution by stagewise recursion, beginning at node $(0, \mathbf{0})$. Figure 1 shows the network representation for the 3×3 table discussed in Section 3.1 having row margins $(6, 1, 2)$ and column margins $(1, 2, 6)$. The topmost path gives the table $(1, 2, 3/0, 0, 1/0, 0, 2)$.

For simply calculating p-values, one can dramatically increase the speed of the network algorithm by computing at each node lower and upper bounds on test statistic values for tables having path passing through that node. In this way, one can determine tables that necessarily do or do not contribute to the p-value, without processing the remaining parts of paths in the network passing through that node. Vollset, Hirji and Elashoff (1991) gave an adaptation of the network algorithm that uses an algebraic rather than geometric representation, treating convolutions of distributions using polynomial multiplication. Their algorithm computes the null distribution on the natural rather than logarithmic scale. Using this approach, they obtained faster computation of "exact" confidence limits for the common odds ratio in $2 \times 2 \times K$ tables.

These days, algorithms such as the network algorithm can practically handle analyses for most $I \times J$ and $2 \times 2 \times K$ tables having a small to moderate number of cells and small to moderate cell counts. To illustrate, Table 6 reports the CPU time on a

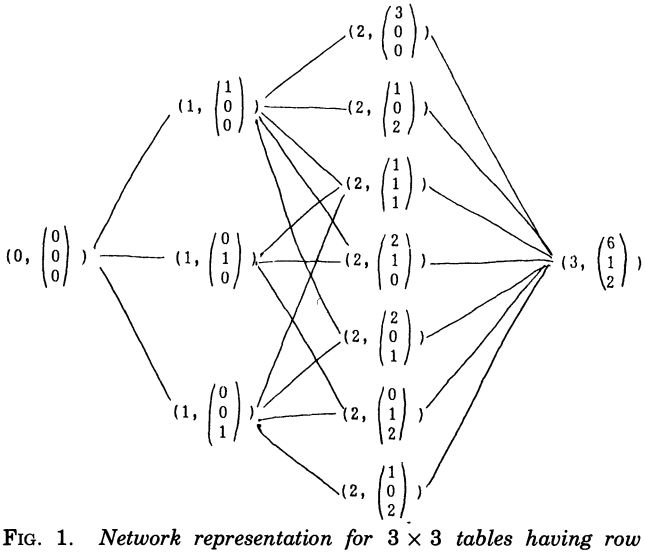

FIG. 1. *Network representation for* 3×3 *tables having row margin* $(6, 1, 2)$ *and column margin* $(1, 2, 6)$.

TABLE 6

Sample CPU times (seconds) for Freeman–Halton test on $I \times I$ *tables with uniform marginal counts and p-values approximately* 0.05, *using SAS on a DEC* 3100 *workstation*

	Total sample size							
I	10	20	30	40	50	60	80	100
2	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
3	< 1	< 1	< 1	< 1	< 1	< 1	1	1
4	< 1	1	2	2	14	22	38	643
5	1	2	12	40	42	72	1037	20,793
6	1	37	107	260	724	8564	[a]	[a]
7	9	118	1270	2123	11,601	[a]	[a]	[a]

[a] More than 6 hours.

DEC 3100 workstation needed for SAS (PROC FREQ) to perform the Freeman–Halton test for $I \times I$ tables of various sizes having uniform marginal totals and cell frequencies changed from uniformity sufficiently to give p-values approximately equal to 0.05. Times are much faster when marginal counts are nonuniform. For instance, a 5×5 table with $n = 100$ having both margins equal to $(60, 25, 10, 3, 2)$ took only 40 seconds CPU time.

The remaining problematic area is large, sparse tables, for which there are a large number of cells and cell counts too small to appeal to standard asymptotic theory. For two-way tables, examples would be tables of at least about 50 cells, having several fitted values less than about 5. Even with current computing power, the conditional reference set for such tables is often too large to be handled by exact methods. Moreover, the cardinality grows so rapidly as a function of the number of cells that, regardless of future improvements in computer power, it may always be possible to produce tables that cannot be handled exactly.

A good compromise for handling large, sparse tables is to estimate precisely the inferential characteristics of interest, such as exact p-values and confidence intervals. One can do this using Monte Carlo sampling of tables in the reference set, by simulating the conditional hypergeometric sampling distribution (Agresti, Wackerly and Boyett, 1979; Boyett, 1979; Cox and Plackett, 1980; Patefield, 1981, 1982; Kreiner, 1987; StatXact, 1991). Each sampled table provides a Bernoulli random variable, indicating whether the test statistic is at least as large as observed. The estimated exact p-value is the sample mean of those Bernoulli random variables, which is the proportion of sampled tables that have test statistics at least as large as the observed one. The precision of the estimate is determined by the estimated sample variance $P(1 - P)/N$, where N is the number of tables sampled. Sampling of 17,000 tables ensures the estimate is good to within 0.01 with confidence at least 0.99.

Agresti (1990, page 308) gave a 16×5 table with 219 observations, relating various characteristics of alligators to their primary food choice (having categories: fish, invertebrate, reptile, bird, other). For such large tables, one could either sample a fixed number of tables N to guarantee a certain accuracy, or repeatedly take samples of N' tables until achieving a certain accuracy. Using the latter approach with $N' = 2{,}000$ and desired accuracy 0.0005 for a p-value for testing independence with the Pearson statistic, Monte Carlo sampling of 10,000 tables provides an estimated exact p-value of 0.0004 and a 99% confidence interval for that exact p-value of $(0.0000, 0.0009)$.

An advantage of the Monte Carlo method is that the amount of computational work is much less dependent on the sample size n and table size $I \times J$ than for methods for exact analysis. For a method of simulating tables that involves taking a random permutation of n integers, Agresti, Wackerly and Boyett (1979) noted that the CPU time is approximately linear in n and stable in I and J. Patefield (1981) provided a method that is more efficient for large n. Although it takes longer to generate each table with Monte Carlo methods, only a relatively small number need to be generated. In principle, this method could be used to approximate precisely any exact analysis, including those for which exact calculations may never be feasible.

Mehta, Patel and Senchaudhuri (1988) described a more sophisticated and faster Monte Carlo approach, using importance sampling. Tables are sampled from the conditional reference set in proportion to their importance for reducing the variance of the estimated p-value, rather than in proportions corresponding to their hypergeometric probabilities. In importance sampling, each sampled table provides an estimate of the p-value that is designed to be much better than the crude Bernoulli estimate provided by Monte Carlo sampling. The tables are sampled using a network algorithm. For linear rank tests for $2 \times J$ tables, they noted that importance sampling can be up to four orders of magnitude more efficient than Monte Carlo sampling. That is, to achieve a certain fixed accuracy with a p-value estimate, the ratio of the number of tables sampled using the Monte Carlo approach versus importance sampling was about 10,000 for tests such as the trend test. However, the initial overhead involved in using importance sampling, due partly to using backward induction with the network algorithm to set up the network-based sampling scheme, makes it inefficient for certain very large data sets.

7.2 Software

Until recently, software for exact methods for contingency tables was nearly nonexistent, at least in the most popular statistical packages. Even now, nearly all packages can perform Fisher's exact test but little if anything else. With a couple of exceptions, our discussion here is limited to the most commonly used packages.

SAS (using procedure FREQ) and IMSL (using routine CTPRB) can perform Fisher's exact test and the Freeman–Halton extension for $I \times J$ tables, but do not give options to perform tests that base the p-values on goodness-of-fit statistics or

ordinal statistics. SAS uses the network algorithm from Mehta and Patel (1983), whereas IMSL uses an algorithm that enumerates the entire reference set. Thus, although neither program seems to have limits on table sizes, SAS can handle a much greater variety of tables in a reasonable amount of time. Currently, BMDP and SPSS^X only perform Fisher's exact test. All these packages seem to use the table probability as the basis of ordering the reference set for two-sided p-values.

StatXact (1991) is a statistical package specializing in exact nonparametric inference and in exact inference for contingency table problems. Developed by Mehta and Patel and colleagues, it uses versions of the network algorithm described in their articles. For $2 \times J$ tables, StatXact performs a general linear rank test that includes as special cases the Wilcoxon test and a trend test with arbitrary scores. For $I \times J$ tables with $\min(I, J) \leq 5$, it can perform the Freeman–Halton test and exact tests of independence using the Pearson or likelihood-ratio chi-squared statistics. For tables with ordered columns, it can perform the exact test using the Kruskal–Wallis statistic when $\max(I, J) \leq 5$. When rows are also ordered, it can perform the exact test of linear-by-linear association described by Agresti, Mehta and Patel (1990) when $\min(I, J) \leq 5$, and it can use the Jonckheere–Terpstra statistic when $I \leq 5$. For $2 \times 2 \times K$ tables, it performs Birch's exact test of conditional independence (for $K \leq 200$), Zelen's exact test of homogeneity of odds ratios, and "exact" confidence intervals for an assumed common odds ratio. For $2 \times J \times K$ tables, StatXact performs stratified linear rank tests and can perform inference for parameters in conditional logistic regression models. For $I \times J$ and $2 \times J \times K$ tables, StatXact performs Monte Carlo sampling for cases beyond its capability for exact inference (as long as $I \leq 50$, $J \leq 50$, $K \leq 200$, and $IJK \leq 2500$), and it can perform importance sampling for $2 \times J$ tables. The StatXact manual is also a good reference for many examples of exact conditional analyses. Many of the StatXact routines are also available in the EGRET statistical software package (EGRET, 1991).

Baptista and Pike (1977) gave a FORTRAN program for the Sterne-type confidence interval for the odds ratio in a single 2×2 table. For $2 \times 2 \times K$ tables, Thomas (1975) gave a FORTRAN program for the conditional ML estimate and an "exact" confidence interval for a common odds ratio, and for exact tests of conditional independence and no three-factor interaction (when $K \leq 20$). This program can be slow since, unlike algorithms presented by Pagano and Tritchler (1983b) and StatXact (1991), it requires evaluating every table in the conditional reference set. Vollset and Hirji (1991) gave a fast GAUSS program for the exact test of conditional independence and confidence interval for a common odds ratio, and indicated that it can handle up to about 1,000 points in the distribution of $\sum_k n_{11k}$.

8. OTHER APPROACHES TO EXACT INFERENCE

This discussion has focused on the classical conditional approach to exact inference for contingency tables. This section discusses controversies regarding that approach and describes alternative approaches that produce results having some connection with those for exact conditional methods.

8.1 Controversy Over Exact Conditional Approaches

Most of the debate about exact conditional methods for categorical data has focused on their use with Fisher's exact test when both margins of the table are not naturally fixed. I discuss the controversy only briefly, as it has already generated an enormous literature. See, for instance, Barnard (1945, 1947, 1949, 1979, 1989), Berkson (1978), Basu (1979), Kempthorne (1979), Upton (1982), Suissa and Shuster (1984, 1985), Yates and discussants (1984), Bhapkar (1986), Haber (1987, 1989), D'Agostino, Chase and Belanger (1988), Lloyd (1988b), Rice (1988), Little (1989), Camilli (1990), Mehta and Hilton (1990), Routledge (1990), Storer and Kim (1990), and Greenland (1991).

The perceived problem with the test results mainly from the conditional distribution of n_{11} (or the odds ratio) being highly discrete, much more so than when one or neither margin is fixed. This results in the test being quite conservative in a conditional or unconditional sense, when used with a fixed significance level such as $\alpha = 0.05$. The actual probability of rejecting the null hypothesis may be considerably less than the nominal level. Proponents of Fisher's test (e.g., Yates, 1984) argue that (1) one should not use arbitrary fixed significance levels (versus simply reporting the p-value), (2) one should not average into the calcu lation of the p-value other tables whose marginals did not occur, and (3) no substantive loss of information about H_0 results from conditioning on the marginals (i.e., the marginal counts are approximately ancillary).

The randomized-decision version of Fisher's exact test, using randomization on the boundary of a critical region to achieve a fixed significance level α, is UMPU (Tocher, 1950). This is of little solace for practical work, since randomization is not used, although it indicates that Fisher's test may also

perform well for the mid-p definition of a p-value. The mid-p value is half the probability of the observed result plus the probability of more extreme values. For discrete data, the mid-p value has null behavior more nearly like a uniform (0, 1) random variable than the ordinary p-value (e.g., its null expected value is 0.5, and the sum of its two one-tailed p-values equals 1.0). It has been recommended (e.g., by Lancaster, 1961; Plackett in the discussion of Yates, 1984; Barnard, 1989, 1990) as a good compromise between having a conservative test and using randomization on the boundary to eliminate problems from discreteness. For comparing two binomial probabilities π_1 and π_2, Hirji, Tan and Elashoff (1991) noted that the mid-p adjustment to Fisher's exact test has actual levels of significance closer to nominal levels than do classical asymptotic tests. This is especially true when the common value of $\{\pi_i\}$ is near 0 or 1 or when the sample sizes n_{1+} and n_{2+} are quite different. Haber (1986b) obtained similar conclusions both for comparing binomials and for multinomial sampling over the four cells. Hirji (1991) showed that, for tests of parameters in conditional logistic models for case-control designs with unmatched binary covariates, mid-p adjustments to an exact test perform well in approximating nominal levels compared to the exact test and asymptotic score tests.

Analogous remarks apply to "exact" interval estimation. Although necessarily conservative, "exact" interval estimates for the odds ratio can be so highly conservative as to be less useful than asymptotic large-sample approaches in terms of long-run coverage performance. An adaptation of Cornfield's "exact" method uses a mid-p adjustment, choosing θ endpoints that have one-sided mid-p values of $\alpha/2$. This approach does not guarantee the desired coverage probability, but simulations by Mehta and Walsh (1992) showed that it performs well in this respect. For small samples, the mid-p-adjusted intervals can be much narrower. For instance, for Fisher's tea-tasting data, having rows (3, 1/1, 3), the "exact" 95% confidence interval is (0.21, 626.2), and the mid-p-based 95% confidence interval is (0.31, 308.6). Vollset, Hirji and Afifi (1991) showed similar good performance of the mid-p approach for interval estimation of parameters in conditional logistic designs.

8.2 Exact Unconditional Approach

Other methods for 2×2 tables have been motivated by the controversy about the conditioning argument and the conservativeness of Fisher's exact test. Some authors argue that it is better to use a robust asymptotic test than a possibly highly conservative exact test (e.g., D'Agostino, Chase and Belanger, 1988). Others recommend using an "exact" unconditional test.

To illustrate the latter approach, consider testing $\pi_1 = \pi_2$ ($\theta = 1$) for two independent binomial samples. One first computes an exact p-value $P(\pi)$ for each possible common value π of π_1 and π_2. For this p-value, one might use the product binomial probability that a z statistic (or chi-squared statistic) for comparing two proportions is at least as large as observed, when π is the value of the nuisance parameter. The global p-value for the test with unknown π is then $P = \sup_\pi P(\pi)$. Proposed by Barnard (1945, 1947), this test was disavowed by him in later publications (e.g., Barnard, 1949, discussion of Yates, 1984). Boschloo (1970), McDonald, Davis and Milliken (1977), Suissa and Shuster (1984, 1985), Haber (1986a, 1987, 1989), Shuster (1988) discussed computational implementation of the unconditional test. Suissa and Shuster (1991) extended it to comparisons of dependent proportions for matched pairs.

The 2×2 table having entries (3, 0/0, 3), discussed by Barnard (1945) and Fisher (1945) [see also Little (1989) and Routledge (1990)], illustrates that results with the conditional and unconditional approaches can be quite discrepant. For H_1: $\theta \neq 1$, the Fisher p-value is $2\binom{3}{0}\binom{3}{3}\Big/\binom{6}{3} = 0.100$, and the asymptotic Pearson chi-squared test has a p-value of 0.014. For the exact binomial test of $\pi_1 = \pi_2$ having only fixed row totals (3, 3), the Pearson chi-squared value of 6.0 that occurs for the observed table and for table (0, 3/3, 0) is the maximum possible, and the p-value for given nuisance parameter π is $2\pi^3(1-\pi)^3$. The supremum of this over $0 \le \pi \le 1$ occurs at $\pi = 1/2$, giving p-value = 1/32. This unconditional p-value is related to Fisher conditional p-values by

$$P(X^2 \ge 6) = \sum_{k=0}^{6} P(X^2 \ge 6 \mid n_{+1} = k) P(n_{+1} = k).$$

But the conditional probability that $X^2 \ge 6$ is zero except when $n_{+1} = 3$, so the unconditional p-value is a weighted average of the Fisher p-value for the observed column marginals and p-values of 0 corresponding to the impossibility of getting results as extreme as observed if other marginals had occured, that is, $1/32 = 0.10\left[\binom{6}{3}(1/2)^6\right]$, where the term in brackets is the binomial probability (when $\pi = 1/2$) of the column marginals (3, 3). Fisher (1945) remarked, "It is my view that the existence of these less informative possibilities should not affect our judgment of significance based on the series actually observed . . . . The fact that such an unhelpful outcome as these might occur, or must

occur with a certain probability, is surely no reason for enhancing our judgment of significance in cases where it has not occurred; ... it is only the sampling distribution of samples of the same type that can supply a rational test of significance."

The unconditional test is so computationally intensive that it seems complex to extend it to larger contingency-table problems (Mehta and Hilton, 1990). In addition, removing nuisance parameters by taking a supremum may itself produce quite conservative analyses for tables having larger numbers of such parameters. It should be noted that asymptotic procedures can also be quite conservative. For instance, Koehler and Larntz (1980) noted that the likelihood-ratio test tends to be highly conservative when most expected frequencies are smaller than 0.5. To illustrate, consider the 3×9 table (0, 7, 0, 0, 0, 0, 0, 1, 1/1, 1, 1, 1, 1, 1, 1, 0, 0/0, 8, 0, 0, 0, 0, 0, 0, 0), discussed in the StatXact manual. For the likelihood-ratio statistic, the asymptotic p-value is 0.0837 and the exact p-value is 0.0015; for the Pearson statistic, the values are 0.1342 and 0.0013.

8.3 Bayesian Approaches

For 2×2 tables, Bayesian approaches using certain "conservative" prior distributions give results equivalent to conditional tests. For instance, Altham (1969) gave an exact Bayesian analysis comparing parameters for two independent binomial samples. She tested H_0: $\pi_1 \leq \pi_2$ against $\pi_1 > \pi_2$ using a beta(α_i, β_i) prior distribution for π_i; i.e., the prior for π_i is proportional to $(\pi_i)^{\alpha}(1 - \pi_i)^{\beta}$, with $\alpha = \alpha_i - 1$ and $\beta = \beta_i - 1$, $i = 1, 2$. The posterior distributions are beta(α_i', β_i') with $\alpha_i' = \alpha_i + n_{i1}$ and $\beta_i' = \beta_i + n_{i2}$. Taking the Bayesian p-value to be the posterior probability that $\pi_1 \leq \pi_2$, Altham showed this equals the one-sided p-value for Fisher's exact test when one uses improper prior distributions $(\alpha_1, \beta_1) = (1, 0)$ and $(\alpha_2, \beta_2) = (0, 1)$. This represents prior belief favoring the null hypothesis, in effect penalizing oneself against concluding that $\pi_1 > \pi_2$. If $\alpha_i = \beta_i = \gamma$, $i = 1, 2$, where $0 \leq \gamma \leq 1$, Altham showed that the Bayesian p-value is smaller than the Fisher p-value, and the difference between the two is no greater than the null probability of the observed data.

Altham (1971) also gave Bayesian analyses for dependent proportions. For a simple model in which the probability of success is the same for each subject at a given occasion, she again showed that the classical p-value is a Bayesian p-value for a prior distribution favoring the null hypothesis. For a model similar to Cox's, in which the probability of success varies by subject but the occasion effect is constant, she showed that the Bayesian evidence against the null is weaker as the number of pairs giving the same response at both occasions increases, for fixed n_{12} and n_{21}. This result differs, and is perhaps more intuitively pleasing, than the exact conditional ML result. For examples of other Bayesian analyses for 2×2 tables, see Leonard (1975), Chen and Novick (1984) and Nurminen and Mutanen (1987).

9. FUTURE RESEARCH

By the turn of the century, we should see advances in applicability of exact methodology for contingency tables at least comparable to those of the past decade. One does not need a crystal ball to predict that computer speed will continue to increase and algorithms will be further improved, so that tables not now feasible for analysis soon will be. In addition, it is reasonable to expect development of algorithms to handle new types of categorical data, in particular, more complex relationships for larger tables in higher dimensions.

This article has emphasized exact inference for contingency tables in the context of loglinear modeling. I believe that presenting the methods in the context of inference for parameters in models helps to unify a variety of exact conditional methods. It also helps to identify areas in which additional research would yield fruitful results, such as exact inferences for $I \times J \times K$ tables and a general goodness-of-fit test for loglinear models. This section summarizes other avenues for future research.

An important but difficult area for future research is exact analysis of model parameters for contingency tables that are large and sparse. For such data, standard asymptotic methods behave poorly; yet, it has been impractical to apply exact methods. The size of the conditional reference set is often too large to handle when the table has a large number of cells. This can also happen when the table is small but contains very large as well as small cell counts. For such problematic tables, there should be additional research on (1) hybrid algorithms that use exact methods for some parts of the computation and approximations for other parts (Baglivo, Olivier and Pagano, 1988), (2) fast ways of simulating the exact distribution (Kreiner, 1987; Mehta, Patel and Senchaudhuri, 1988), and (3) better asymptotic approximations (e.g., Koehler, 1986), including saddlepoint approximations (Davison, 1988; Booth and Butler, 1990; Bedrick and Hill, 1992; Pierce and Peters, 1992).

An area in which large, sparse tables commonly occur is modeling of longitudinal data with categorical responses. Conditioning on sufficient statistics

to eliminate nuisance parameters, such as subject random effects, can be very helpful. Even then, the large number of cells in the table make exact methods or approximations for them quite challenging. Another complication is that some useful models do not have reductions of data through sufficiency, such as loglinear and logistic models for the marginal probabilities rather than joint cell probabilities.

An area that has scope for lots of additional work is exact inference for $I \times J \times K$ tables. Section 4.4 described several exact inferences for standard loglinear models. Other exact inferences are of interest when at least one variable is ordinal. In testing conditional independence, one could use test statistics designed to improve power for narrower alternatives. For instance, Section 4.4 discussed the possibility of testing conditional independence against a linear-by-linear alternative. When X is nominal and Y is ordinal, one could use a statistic analogous to a stratified Kruskal–Wallis statistic. Such a statistic results from comparing (XZ, YZ) to a loglinear model in which the X-Y partial association term has the form $\mu_i y_j$, for parametric row effects $\{\mu_i\}$ and rank scores $\{y_j\}$. Efficient score statistics for these cases are equivalent to statistics proposed by Birch (1965) and Landis, Heyman and Koch (1978) for large-sample testing of conditional independence. Similar remarks apply to exact tests of no three-factor interaction. Greater power for narrower alternatives (unsaturated models for three-factor interaction) would be achieved by using statistics that treat one, two or all three variables as ordinal.

Of course, exact inferences for standard loglinear models and for models recognizing ordinality are also relevant for four-way and higher dimensional tables. Many tests of mutual independence or conditional independence can be re-expressed in terms of two-way or three-way tables, but not all. Also, complications may result in tests about three-factor and higher order interactions. The reference set of tables having the required fixed marginals may be difficult to generate or simulate. For instance, in a four-way table, an exact test of no three-factor interaction deals with a reference set in which all six two-way margins are fixed.

There is unlikely to be as much controversy with exact conditional methods for models for higher dimensional tables as there has been with Fisher's exact test for 2×2 tables. That controversy is probably due less to philosophical disagreements about conditioning than to practical consequences of using continuously defined measures such as p-values with highly discrete distributions. Except in extreme cases in which nearly all observations fall in one or two marginal categories for each classification, sampling distributions are much "less discrete" for larger tables. As the number of dimensions or the number of response categories per dimension increases, the sampling distributions that determine exact tests and confidence intervals rapidly approach a continuous form. Thus, the conservativeness of the inferences becomes less problematic. For instance, performance of exact conditional tests should be closer to that of their randomized versions, many of which are UMPU. Also, the three approaches discussed for confidence intervals for odds ratios (test-based with $\alpha/2$ in each tail, adapted Sterne, test-based with mid-p values) should all have true confidence levels close to the nominal one.

For cases in which discreteness is problematic, there should be more work on developing adaptations of exact methods that, while perhaps sacrificing "exactness," give performance closer to the randomized exact versions. Examples are adaptations of exact tests and confidence intervals using the mid-p value (e.g., Hirji, Tan and Elashoff, 1991; Mehta and Walsh, 1992), and tests and confidence intervals that use the exact distribution computed at the ML estimate of the nuisance parameter (e.g., Conlon and Thomas, 1990; Storer and Kim, 1990).

So far there has been relatively little attention paid to algorithms for calculating power for exact conditional tests. The literature refers almost exclusively to 2×2 tables. For instance, Gail and Gart (1973), Haseman (1978) and Suissa and Shuster (1985) discussed the determination of sample size for obtaining desired powers in fixed-α use of Fisher's exact test.

Finally, it would be very useful to have a general-purpose algorithm for exact tests comparing two nested loglinear models. This would unify exact tests of goodness of fit for all loglinear models, as well as tests against unsaturated alternatives.

In summary, much has been accomplished in the past decade, but this is only the tip of the iceberg of what can be done. In the future, statistical practice for categorical data should place more emphasis on "exact" methods and less emphasis on methods using possibly inadequate large-sample approximations.

ACKNOWLEDGMENTS

This work was partially supported by NIH Grant GM43824. The author appreciates helpful comments on manuscript organization and content from Dr. Cyrus Mehta and two referees. Computations

for many of the examples presented in this article were obtained using StatXact (1991).

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

AGRESTI, A., MEHTA, C. R. and PATEL, N. R. (1990). Exact inference for contingency tables with ordered categories. *J. Amer. Statist. Assoc.* **85** 453-458.

AGRESTI, A. and WACKERLY, D. (1977). Some exact conditional tests of independence for $R \times C$ cross-classification tables. *Psychometrika* **42** 111-125.

AGRESTI, A., WACKERLY, D. and BOYETT, J. (1979). Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* **44** 75-83.

ALTHAM, P. M. E. (1969). Exact Bayesian analysis of a 2×2 contingency table and Fisher's 'exact' significance test. *J. Roy. Statist. Soc. Ser. B* **31** 261-269.

ALTHAM, P. M. E. (1971). The analysis of matched proportions. *Biometrika* **58** 561-576.

ANDERSEN, A. H. (1974). Multidimensional contingency tables. *Scand. J. Statist.* **1** 115-127.

ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11** 375-386.

BAGLIVO, J., OLIVIER, D. and PAGANO, M. (1988). Methods for the analysis of contingency tables with large and small cell counts. *J. Amer. Statist. Assoc.* **83** 1006-1013.

BAKER, R. J. (1977). Exact distributions derived from two-way tables. *J. Roy. Statist. Soc. Ser. C* **26** 199-206.

BALMER, D. W. (1988). Recursive enumeration of $r \times c$ tables for exact likelihood evaluation. *J. Roy. Statist. Soc. Ser. C* **37** 290-301.

BAPTISTA, J. and PIKE, M. C. (1977). Exact two-sided confidence limits for the odds ratio in a 2×2 table. *J. Roy. Statist. Soc. Ser. C* **26** 214-220.

BARNARD, G. A. (1945). A new test for 2×2 tables (Letter to the Editor). *Nature* **156** 177.

BARNARD, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* **34** 123-138.

BARNARD, G. A. (1949). Statistical inference. *J. Roy. Statist. Soc. Ser. B* **11** 115-139.

BARNARD, G. A. (1979). In contradiction to J. Berkson's dispraise: Conditional tests can be more efficient. *J. Statist. Plann. Inference* **3** 181-188.

BARNARD, G. A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine* **8** 1469-1477.

BARNARD, G. A. (1990). Must clinical trials be large? The interpretation of P-values and the combination of test results. *Statistics in Medicine* **9** 601-614.

BASU, D. (1979). Discussion of "In dispraise of the exact test" by J. Berkson. *J. Statist. Plann. Inference* **3** 189-192.

BAYER, L. and COX, C. (1979). Exact tests of significance in binary regression models. *J. Roy. Statist. Soc. Ser. C* **28** 319-324.

BEDRICK, E. J. and HILL, J. R. (1990). Outlier tests for logistic regression: A conditional approach. *Biometrika* **77** 815-827.

BEDRICK, E. J. and HILL, J. R. (1992). An empirical assessment of saddlepoint approximations for testing a logistic regression parameter. *Biometrics*. To appear.

BERKSON, J. (1978). In dispraise of the exact test. *J. Statist. Plann. Inference* **2** 27-42.

BHAPKAR, V. P. (1986). On conditionality and likelihood with nuisance parameters in models for contingency tables. Technical Report 253, Dept. Statistics, Univ. Kentucky.

BIRCH, M. W. (1964). The detection of partial association I: The 2×2 case. *J. Roy. Statist. Soc. Ser. B* **26** 313-324.

BIRCH, M. W. (1965). The detection of partial association II: The general case. *J. Roy. Statist. Soc. Ser. B* **27** 111-124.

BISHOP, Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics* **27** 545-562.

BOOTH, J. and BUTLER, R. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77** 787-796.

BOSCHLOO, R. D. (1970). Raised conditional level of significance for the 2×2 table when testing for the equality of two probabilities. *Statist. Neerlandica* **21** 1-35.

BOULTON, D. M. (1974). Remark on algorithm 434. *Comm. ACM* **17** 326.

BOYETT, J. (1979). Random $R \times C$ tables with given row and column totals. *J. Roy. Statist. Soc. Ser. C* **28** 329-332.

BRESLOW, N. E. and DAY, N. E. (1980). *The Analysis of Case-Control Studies*. IARC Scientific Publications No. 32, Lyon, France.

CAMILLI, G. (1990). The test of homgeneity for 2×2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin* **108** 135-145.

CANTOR, A. B. (1979). A computer algorithm for testing significance in $M \times K$ contingency tables. In *Proceedings of the Statistical Computing Section* 220-221. Amer. Statist. Assoc., Washington, DC.

CHEN, J. J. and NOVICK, M. R. (1984). Bayesian analysis for binomial models with general beta prior distributions. *Journal of Educational Statistics* **9** 163-175.

COCHRAN, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics* **10** 417-451.

COHEN, A. and SACKROWITZ, H. B. (1991). Tests for independence in contingency tables with ordered alternatives. *J. Multivariate Anal.* **36** 56-67.

CONLON, M. and THOMAS, R. G. (1990). A new confidence interval for the difference of two binomial proportions. *Comput. Statist. Data Anal.* **8** 237-241.

COOLEY, J. M. and TUKEY, J. W. (1965). An algorithm for the machine calculation of complex Fourier transforms. *Math. Comp.* **12** 297-301.

CORNFIELD, J. (1956). A statistical problem arising from retrospective studies. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **4** 135-148 Univ. California Press, Berkeley.

COX, D. R. (1958a). Two further applications of a model for binary regression. *Biometrika* **45** 562-565.

COX, D. R. (1958b). The regression analysis of binary sequences (with discussion). *J. Roy. Statist. Soc. Ser. B* **20** 215-242.

COX, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika* **53** 215-220.

COX, D. R. (1970). *Analysis of Binary Data*. Chapman and Hall, London.

COX, M. A. A. and PLACKETT, R. L. (1980). Small samples in contingency tables. *Biometrika* **67** 1-13.

CRESSIE, N. and READ, T. R. C. (1989). Pearson X^2 and the loglikelihood ratio statistic G^2: A comparative review. *Internat. Statist. Rev.* **57** 19-43.

D'AGOSTINO, R. B., CHASE, W. and BELANGER, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Amer. Statist.* **42** 198-202.

DAVIS, L. J. (1986). Exact tests for 2 by 2 contingency tables. *Amer. Statist.* **40** 139-141.

DAVISON, A. C. (1988). Approximate conditional inference in generalized linear models. *J. Roy. Statist. Soc. Ser. B* **50** 445-461.

DUPONT, W. D. (1986). Sensitivity of Fisher's exact test to minor perturbations in 2×2 contingency tables. *Statistics in Medicine* **5** 629-635.

EGRET. (1991). *EGRET Statistical Software*. Statistics and Epidemiology Research Corporation, Seattle.

FISHER, R. A. (1934). *Statistical Methods for Research Workers*. (Originally published 1925, 14th ed. 1970.) Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935a). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc. Ser. A* **98** 39-82.

FISHER, R. A. (1935b). *The Design of Experiments* (8th ed. 1966). Oliver and Boyd, Edinburgh.

FISHER, R. A. (1945). A new test for 2 × 2 tables (Letter to the Editor). *Nature* **156** 388.

FISHER, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics* **6** 17-24.

FREEMAN, G. H. and HALTON, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* **38** 141-149.

GAIL, M. H. and GART, J. J. (1973). The determination of sample sizes for use with the exact conditional test in 2 × 2 comparative trials. *Biometrics* **29** 441-448.

GAIL, M. H. and MANTEL, N. (1977). Counting the number of $r \times c$ contingency tables with fixed margins. *J. Amer. Statist. Assoc.* **72** 859-862.

GART, J. J. (1969). An exact test for comparing matched proportions in crossover designs. *Biometrika* **56** 75-80.

GART, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. *Biometrika* **57** 471-475.

GASTWIRTH, J. L. (1988). *Statistical Reasoning in Law and Public Policy* 1. Academic, San Diego.

GIBBONS, J. D. and PRATT, J. W. (1975). *P*-values: Interpretation and methodology. *Amer. Statist.* **29** 20-25.

GOOD, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4** 1159-1189.

GOOD, I. J. (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Math.* **19** 23-45.

GOOD, I. J. (1982). The fast calculation of the exact distribution of Pearson's chi-squared and of the number of repeats within the cells of a multinomial by using a Fast Fourier Transform. *J. Statist. Comput. Simulation* **14** 71-78.

GOOD, I. J., GOVER, T. N. and MITCHELL, G. J. (1970). Exact distributions for X^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *J. Amer. Statist. Assoc.* **65** 267-283.

GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** 537-552.

GRAUBARD, B. I. and KORN, E. L. (1987). Choice of column scores for testing independence in ordered 2 × *K* tables. *Biometrics* **43** 471-476.

GREENLAND, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* **45** 248-251.

HABER, M. (1986a). An exact unconditional test for the 2 × 2 comparative trial. *Psychological Bulletin* **99** 129-132.

HABER, M. (1986b). A modified exact test for 2 × 2 contingency tables. *Biometrical J.* **28** 455-463.

HABER, M. (1987). A comparison of some conditional and unconditional exact tests for 2 by 2 contingency tables. *Comm. Statist. Simulation Comput.* **16** 999-1013.

HABER, M. (1989). Do the marginal totals of a 2 × 2 contingency table contain information regarding the table proportions? *Comm. Statist. Theory Methods* **18** 147-156.

HABERMAN, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics* **36** 589-600.

HABERMAN, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5** 1148-1169.

HABERMAN, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.* **83** 555-560.

HASEMAN, J. K. (1978). Exact sample sizes for use with the Fisher-Irwin test for 2 × 2 tables. *Biometrics* **34** 106-110.

HEALY, M. J. R. (1969). Exact tests of significance in contingency tables. *Technometrics* **11** 393-395.

HILTON, J. F., MEHTA, C. R. and PATEL, N. R. (1991). Exact Smirnov tests using a network algorithm. Technical Report 14, Dept. Epidemiology and Biostatistics, Univ. California, San Francisco.

HIRJI, K. F. (1991). A comparison of exact, mid-*P*, and score tests for matched case-control studies. *Biometrics* **47** 487-496.

HIRJI, K. F., MEHTA, C. R. and PATEL, N. R. (1987). Computing distributions for exact logistic regression. *J. Amer. Statist. Assoc.* **82** 1110-1117.

HIRJI, K. F., MEHTA, C. R. and PATEL, N. R. (1988). Exact inference for matched case-control studies. *Biometrics* **44** 803-814.

HIRJI, K. F., TAN, S. and ELASHOFF, R. M. (1991). A quasi-exact test for comparing two binomial parameters. *Statistics in Medicine* **10** 1137-1153.

HIRJI, K. F., TSIATIS, A. A. and MEHTA, C. R. (1989). Median unbiased estimation for binary data. *Amer. Statist.* **43** 7-11.

IRWIN, J. O. (1935). Tests of significance for differences between percentages based on small numbers. *Metron* **12** 83-94.

KEMPTHORNE, O. (1979). In dispraise of the exact test: Reactions. *J. Statist. Plann. Inference* **3** 199-213.

KLOTZ, J. (1966). The Wilcoxon, ties, and the computer. *J. Amer. Statist. Assoc.* **61** 772-787.

KLOTZ, J. and TENG, J. (1977). One-way layout for counts and the exact eumeration of the Kruskal-Wallis *H* distribution with ties. *J. Amer. Statist. Assoc.* **72** 165-169.

KOEHLER, K. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* **81** 483-493.

KOEHLER, K. and LARNTZ, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* **75** 336-344.

KREINER, S. (1987). Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scand. J. Statist.* **14** 97-112.

KURITZ, S. J., LANDIS, J. R. and KOCH, G. G. (1988). A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health* **9** 123-160.

LANCASTER, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56** 223-234.

LANDIS, J. R., HEYMAN, E. R. and KOCH, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Internat. Statist. Rev.* **46** 237-254.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **37** 23-37.

LITTLE, R. J. A. (1989). Testing the equality of two independent binomial proportions. *Amer. Statist.* **43** 283-288.

LLOYD, C. J. (1988a). Doubling the one-sided *P*-value in testing independence in 2 × 2 tables against a two-sided alternative. *Statistics in Medicine* **7** 1297-1306.

LLOYD, C. J. (1988b). Some issues arising from the analysis of 2 × 2 contingency tables. *Austral. J. Statist.* **30** 35-46.

MANTEL, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Amer. Statist. Assoc.* **58** 690-700.

MANTEL, N. (1987). Exact tests for 2×2 contingency tables (Letter). *Amer. Statist.* **41** 159.

MANTEL, N. and HANKEY, B. J. (1971). Programmed analysis of a 2×2 contingency table. *Amer. Statist.* **25** 40-44.

MARCH, D. L. (1972). Exact probabilities for $R \times C$ contingency tables. *Comm. ACM* **15** 991-992.

MCCULLAGH, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81** 104-107.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MCDONALD, L. L., DAVIS, B. M. and MILLIKEN, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables. *Technometrics* **19** 145-158.

MCNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** 153-157.

MEHTA, C. and HILTON, J. (1990). Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. Unpublished manuscript.

MEHTA, C. R. and PATEL, N. R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **78** 427-434.

MEHTA, C. R. and PATEL, N. R. (1986). FEXACT: A Fortran subroutine for Fisher's exact test in unordered $r \times c$ contingency tables. *ACM Trans. Math. Software* **12** 154-161.

MEHTA, C. R. and WALSH, S. J. (1992). Comparison of exact, mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *Amer. Statist.* To appear.

MEHTA, C. R., PATEL, N. R. and GRAY, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Amer. Statist. Assoc.* **80** 969-973.

MEHTA, C. R., PATEL, N. R. and SENCHAUDHURI, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Statist. Assoc.* **83** 999-1005.

MEHTA, C. R., PATEL, N. R. and SENCHAUDHURI, P. (1991). Exact stratified linear rank tests for binary data. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.* (E. M. Keramidas, ed.) 200-207. Interface Foundation, Fairfax Station, VA.

MEHTA, C. R., PATEL, N. R. and TSIATIS, A. A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40** 819-825.

MORRIS, C. N. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165-188.

NEYMAN, J. (1935). On the problem of confidence limits. *Ann. Math. Statist.* **6** 111-116.

NURMINEN, M. and MUTANEN, P. (1987). Exact Bayesian analysis of two proportions. *Scand. J. Statist.* **14** 67-77.

PAGANO, M. and HALVORSEN, K. T. (1981). An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **76** 931-934.

PAGANO, M. and TRITCHLER, D. (1983a). On obtaining permutation distributions in polynomial time. *J. Amer. Statist. Assoc.* **78** 435-440.

PAGANO, M. and TRITCHLER, D. (1983b). Algorithms for the analysis of several 2×2 contingency tables. *SIAM J. Sci. Statist. Comput.* **4** 302-309.

PATEFIELD, W. M. (1981). An efficient method of generating random $R \times C$ tables with given row and column totals. *J. Roy. Statist. Soc. Ser. C* **30** 91-97.

PATEFIELD, W. M. (1982). Exact tests for trends in ordered contingency tables. *J. Roy. Statist. Soc. Ser. C* **31** 32-43.

PEARSON, E. S. (1990). *'Student' A Statistical Biography of William Sealy Gosset* (R. L. Plackett and G. A. Barnard, eds.). Clarendon Press, Oxford, England.

PERITZ, E. (1982). Exact tests for matched pairs: Studies with covariates. *Comm. Statist. Theory Methods* **11** 2165-2166. [Correction: (1983) **12** 1209-1210.]

PIERCE, D. A. and PETERS, D. (1991). Practical use of higher-order asymptotics for multiparameter exponential families. *J. Roy. Statist. Soc.* To appear.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

RICE, W. R. (1988). A new probability model for determining exact P-values for 2×2 contingency tables when comparing binomial proportions. *Biometrics* **44** 1-22.

ROUTLEDGE, R. D. (1990). Resolving the controversy over Fisher's exact test. Paper presented at the International Biometric Conference, Budapest.

SANTNER, T. J. and SNELL, M. K. (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *J. Amer. Statist. Assoc.* **75** 386-394.

SHAPIRO, S., SLONE, D., ROSENBERG, L., KAUFMAN, D., STOLLEY, P. D. and MIETTINEN, O. S. (1979). Oral contraceptive use in relation to myocardial infarction. *Lancet* **8119** 743-746.

SHUSTER, J. J. (1988). EXACTB and CONF: Exact unconditional procedures for binomial data. *Amer. Statist.* **42** 234.

SOMS, A. P. (1985). Permutation tests for k-sample binomial data with comparisons of exact and approximate P-levels. *Comm. Statist. Theory Methods* **14** 217-233.

STATXACT. (1991). *StatXact: Statistical Software for Exact Nonparametric Inference*, Version 2. Cytel Software, Cambridge, Mass.

STERNE, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41** 275-278.

STORER, B. E. and KIM, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *J. Amer. Statist. Assoc.* **85** 146-155.

STUMPF, R. H. and STEYN, H. S. (1986). Exact distributions associated with an $I \times J \times K$ contingency table. *Comm. Statist. Theory Methods* **15** 1889-1904.

SUISSA, S. and SHUSTER, J. J. (1984). Are uniformly most powerful unbiased tests really best? *Amer. Statist.* **38** 204-206.

SUISSA, S. and SHUSTER, J. J. (1985). Exact unconditional samples sizes for the 2 by 2 binomial trial. *J. Roy. Statist. Soc. Ser. A* **148** 317-327.

SUISSA, S. and SHUSTER, J. J. (1991). The 2×2 matched pair trial: Exact unconditional design and analysis. *Biometrics* **47** 361-372.

THOMAS, D. G. (1971). Exact confidence limits for the odds ratio in a 2×2 table. *J. Roy. Statist. Soc. Ser. C* **20** 105-110.

THOMAS, D. G. (1975). Exact and asymptotic methods for the combination of 2×2 tables. *Computers and Biomedical Research* **8** 423-446.

TOCHER, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* **37** 130-144.

TRITCHLER, D. (1984). An algorithm for exact logistic regression. *J. Amer. Statist. Assoc.* **79** 709-711.

UPTON, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *J. Roy. Statist. Soc. Ser. A* **145** 86-105.

VERBEEK, A. and KROONENBERG, P. M. (1985). A survey of algorithms for exact distributions of test statistics in $r \times c$

contingency tables with fixed margins. *Comput. Statist. Data Anal.* **3** 159-185.

VOLLSET, S. E. and HIRJI, K. F. (1991). A microcomputer program for exact and asymptotic analysis of several 2×2 tables. *Epidemiology* **2** 217-220.

VOLLSET, S. E., HIRJI, K. F. and AFIFI, A. A. (1991). Evaluation of exact and asymptotic interval estimators in logistic analysis of matched case-control studies. *Biometrics* **47** 1311-1325.

VOLLSET, S. E., HIRJI, K. F. and ELASHOFF, R. M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *J. Amer. Statist. Assoc.* **86** 404-409.

WHITE, A. A., LANDIS, J. R. and COOPER, M. M. (1982). A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Internat. Statist. Rev.* **50** 27-34.

YATES, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. Roy. Statist. Soc. Supp.* **1** 217-235.

YATES, F. (1984). Tests of significance for 2×2 contingency tables. *J. Royal Statist. Soc. Ser. A* **147** 426-463.

ZELEN, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika* **58** 129-137.

ZELEN, M. (1972). Exact significance tests for contingency tables embedded in a 2^n classification. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 737-757. Univ. California Press, Berkeley.

ZELTERMAN, D. (1987). Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Assoc.* **82** 624-629.

Comment

Edward J. Bedrick and Joe R. Hill

We congratulate Professor Agresti for his comprehensive review of exact inference with categorical data. We share his enthusiasm for exact conditional methods and believe that the coming years will produce many important computational breakthroughs in this area.

The mechanics of conditioning on sufficient statistics to generate reference distributions for estimation, testing and model checking with loglinear models for Poisson data and logistic regression models for binomial data are well-known, but the utility of conditioning in these settings is not universally agreed upon. Furthermore, the role of conditioning in the analysis of discrete generalized linear models with noncanonical link functions has received little attention from most of the statistical community. As a result, scientists and statisticians are familiar with conditional methods, but many are unsure how such methods should be incorporated into an overall strategy for analyzing categorical data. We feel that the use and abuse of conditional methods will not be fully understood or appreciated without such a strategy. We hope that Professor Agresti's survey and the ensuing discussions stimulate further work in this direction.

Edward J. Bedrick is an Assistant Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131. Joe R. Hill is an R&D Specialist at EDS Research, 5951 Jefferson Street, NE, Albuquerque, New Mexico 87109.

CHECKING LOGISTIC REGRESSION MODELS

We would like to convey some of our recent work on model checking for logistic regression and some of our thoughts regarding conditional inference. For the sake of simplicity, we assume that a single model is under consideration. A little notation is required. The usual logistic regression model has two distinct parts: a sampling component and a structural component. The sampling component specifies that $Y = (Y_1, \ldots, Y_n)'$ is a vector of independent binomial random variables with $Y_i \sim \text{Bin}(m_i, \pi_i)$. The structural or regression component of the model is given by

$$\text{logit}(\pi) = X\beta, \tag{1}$$

where $\text{logit}(\pi)$ is an $n \times 1$ vector of log-odds with elements $\log\{\pi_i/(1-\pi_i)\}$, X is an $n \times p$ full-column rank design matrix with ith row x_i', and β is a $p \times 1$ vector of unknown regression parameters. Under model (1), $S = X'Y$ is sufficient for β. Let $\hat{\pi}$ be the MLE of π under this model.

The distribution of the data $\text{pr}(Y; \beta)$, indexed by β, can be factored into the marginal distribution of the sufficient statistic S, and the conditional distribution of the data given the sufficient statistic:

$$\text{pr}(Y; \beta) = \text{pr}(Y \mid S)\text{pr}(S; \beta).$$

Taking a Fisherian point of view (Fisher, 1950), inferences about β are based on $\text{pr}(S; \beta)$, whereas model checks use the conditional distribution $\text{pr}(Y \mid S)$. Letting $s_{obs} = X'y_{obs}$ be the observed value of the sufficient statistic for the logistic model,